AFRL-RH-BR-TR-2008-0050

THE IDENTIFICATION OF FATIGUE RESISTANT AND FATIGUE SUSCEPTIBLE INDIVIDUALS

Richard Harrison
Scott Chaiken
Donald Harville

Air Force Research Laboratory

Joseph Fischer

General Dynamics

Dion Fisher
Jeff Whitmore
Air Force Research Laboratory

May 2008

Air Force Research Laboratory
Human Effectiveness Directorate
Biosciences and Protection Division
Biobehavioral Performance Branch
Brooks City-Base TX

NOTICE AND SIGNATURE PAGE

Using Government drawings, specifications, or other data included in this document for any purpose other than Government procurement does not in any way obligate the U.S. Government. The fact that the Government formulated or supplied the drawings, specifications, or other data does not license the holder or any other person or corporation; or convey any rights or permission to manufacture, use, or sell any patented invention that may relate to them.

This report was cleared for public release by the Air Force Research Laboratory Human Effectiveness Directorate Public Affairs Office and is available to the general public, including foreign nationals. Copies may be obtained from the Defense Technical Information Center (DTIC) (http://www.dtic.mil).

AFRL-RH-BR-TR-2008-0050 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

//SIGNED//
SCOTT CHAIKEN
Contract Monitor

//SIGNED//
MARK M. HOFFMAN
Deputy Division Chief

This report is published in the interest of scientific and technical information exchange, and its publication does not constitute the Government's approval or disapproval of its ideas or findings.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED** *(From - To)*
5-5-2008	Interim Technical Report	May 2006 – June 2007

4. TITLE AND SUBTITLE
The Identification of Fatigue Resistant and Fatigue Susceptible Individuals

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Richard Harrison,[⌂] Scott Chaiken,[⌂] Donald Harville,[⌂] Joseph Fischer,* Dion Fisher[⌂] and Jeff Whitmore[⌂]

5d. PROJECT NUMBER
7757

5e. TASK NUMBER
P9

5f. WORK UNIT NUMBER
18

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

⌂Air Force Material Command Brooks City-Base, TX
Air Force Research Laboratory 78235
Human Effectiveness Directorate
Biosciences and Protection Division
Biobehavioral Performance Branch
2485 Gillingham Drive

*General Dynamics
Advanced Information Services
5200 Springfield Pike
Dayton, OH 45431

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Air Force Materiel Command
Air Force Research Laboratory
Human Effectiveness Directorate
Biosciences and Protection Division
Biobehavioral Performance Branch
2485 Gillingham Drive
Brooks City-Base, TX 78235

10. SPONSOR/MONITOR'S ACRONYM(S)
AFRL/RH, AFRL/RHP,

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
AFRL-RH-BR-TR-2008-0050

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

The present study was designed to target two specific areas regarding fatigue. The primary purpose was to begin investigations into possible genetic markers linked to fatigue resistance and fatigue susceptibility. This study provided a resistant or susceptible listing of individuals to a genetics research group that is correlating the rankings to the participant's genetic material. The secondary focus was to follow up past research in order to quantify fatigue's effects on a team performing a complex Command, Control, and Communications (C3) type task. Participants completed two, four-hour training sessions prior to experiencing a 48-hr period of sustained wakefulness. During the testing session participants iteratively took part in various cognitive performance tasks as well as a complex air battle management task (either as an individual or team depending on their assignment). At the end of the 48 hours, performance on all measures showed significant effects of fatigue. In order to determine which participants were fatigue resistant/susceptible, a percent change score was used for the various cognitive tasks in order to rank the participants. The lower the percent change, the more resistant a participant was to fatigue on that task. Participants' rankings were then averaged across all of the cognitive tasks in order to produce an overall ranking. When this list is correlated to demographics, the amount of weekday sleep a participant receives significantly impacts the results (r (90) =.36). To remove this potential confounding factor, a second ranking of resistance/susceptibility was created that took into account the amount of sleep the participants reported during the week. In addition to the two fatigue resistant/susceptibility lists, the study found that team productivity was about the same as individual productivity on the complex air battle management task. Also, performance on the complex air battle management tasks (regardless of being a team or individual) degraded less than the conventional cognitive tests. All results are discussed.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
					Lt Richard Harrison
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**	SAR	56	**19b. TELEPHONE NUMBER** *(include area code)*
Unclassified	Unclassified	Unclassified			

This page left intentionally blank

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Fatigue remains an important concern in today's military operations. The present study was designed to target two specific areas regarding fatigue. The primary goal was to begin investigations into possible genetic markers linked to fatigue resistance and fatigue susceptibility. A second goal was to compare individual and team performance when utilizing a complex air battle management task with simple cognitive task performance. Participants completed two, four-hour training sessions prior to experiencing a 48-hr period of sustained wakefulness. During the testing session participants iteratively took part in various cognitive performance tasks as well as the complex air battle management task (either as an individual or team depending on their assignment). At the end of the 48 hours, performance on all measures showed significant effects of fatigue. To determine which participants were fatigue resistant/susceptible, percent change score for the various cognitive tasks to rank the participants was used. The lower the percent change, the more resistant a participant was to fatigue on that task. The participants' rankings across all of the cognitive tasks was then averaged in order to produce an overall ranking. When this list was correlated to demographics, the amount of weekday sleep a participant received significantly impacted the results ($r(90) = .36$). To remove this potential confounding factor, a second ranking of resistance/susceptibility was created that took into account the amount of sleep the participants reported during the week. With regard to the second goal, the study found that team productivity was about the same as individual productivity on the complex air battle management task. Also, performance on the complex air battle management tasks (regardless of being a team or individual) degraded later from fatigue than the conventional cognitive tasks.

INTRODUCTION

One of the most heavily researched human performance topics is sleep deprivation. It strongly impairs human performance (Harrison & Horne, 2000; Pilcher & Huffcutt, 1996), impairs individual performance in a complex manner, and alters performance on some types of tasks much more than others (Lieberman, Tharion, Shukitt-Hale, Speckman, & Tully, 2002). Fatigue insidiously depletes combat effectiveness and elevates the risk to individuals and teams. United States Air Force (USAF) operations are longer in duration, conducted in inhospitable environments, and are counter to normal human diurnal cycles. Commanders currently have no objective real-time measures of individual fatigue status, the capability to identify individuals that are fatigue susceptible or fatigue resistant.

The exact purpose of sleep is unknown. Siegel (2005) defines sleep as a state of immobility with greatly reduced responsiveness. As opposed to a coma or anesthesia, it is characterized by a more rapid reversal to wakefulness. The current report uses the Barnes and Hollenbeck (in press) characterization of sleep deprivation, "as a state of diminished capacity induced by a shortage of sleep." Fatigue resulting from lost or disrupted sleep can occur acutely due to staying awake longer than usual (e.g., for 48 continuous hours), or cumulatively due to having partially restricted sleep over several consecutive days.

Sustained Operations and Fatigue

Sustained operations are a significant and common stress in today's military operations. Combat missions require constant vigilance over time and adaptive performance over prolonged periods. During the early stages of military campaigns members of the command center are often up for several days with little if any time for recuperative sleep. Over time, acute and cumulatively fatigue will affect everyone and the likelihood of error will increase (Hursh, 1998). This is particularly relevant to Command, Control, and Communications (C3) situations, which require constant monitoring for sudden-onset time-critical events.

USAF C3 warfighters face increasingly complex and novel environments with multiple demands for enhanced vigilance, rapid situation assessment, and coordinated adaptive response (Cohen, 1993; Klein, 1993; Mitchell & Beach, 1990; Orasanu & Salas, 1993; Orasanu & Connolly; 1993; Rasmussen, 1993). In tactical C3 situations, the focus is on dynamic battle management and time-critical targeting. Coordination demand is high; reconnaissance and resource allocation depend upon close coordination between ground and air forces in a distributed network system of systems. Situations requiring close coordination and adaptive re-planning are increasingly prevalent and challenging.

It is clear that challenges within these battle scenarios are critically important to air and ground superiority. Much effort has been focused on the development of advanced technology to provide and represent time-critical information during mission execution. These capabilities are needed to facilitate, and even to enable, situation

awareness and coordinated response in conditions of information complexity and time pressure. However, technology can only support, not replace, the role of the war fighters. In fact, it can easily be argued that technology sometimes (perhaps often) increases the role and demands of the human decision maker.

C3 operators must face these ever-increasing demands in conditions that are not ideal. It is, and will be, quite likely that operators will be chronically tired and sleep-deprived. Such is the nature of battlefield operations. During the early stages of actual operations, members of the command center are often up for several days. Over time, chronic fatigue will affect everyone, and the likelihood of error will increase (Bonnet, 2000; Hursh, 1998). This is particularly relevant to C3 situations, which require constant monitoring, even when events are quiet.

Extensive data are available on the negative acute and chronic effects of sleep loss on physiological, attitudinal, and cognitive function (Kryger, Roth, & Dement, 2000). In a review of findings, Bonnett (2000) report an array of negative effects. These effects include mood changes, disorientation, irritability, perceptual distortions, hallucinations, difficulty in concentration, and/or paranoid thinking, depending on the extent of sleep loss. Negative effects have also been demonstrated on a range of cognitive tasks, such as monitoring tasks, speed/accuracy tasks, short-term memory, logical reasoning, and mental subtraction/addition. Physiological effects to fatigue include nystagmus, hand tremor, slurring of speech, sluggish corneal reflexes, hyperactive gag reflex, and increased sensitivity to pain. Unfortunately there is documentation of increases in sleep deprivation in modern society. According to the National Institute for Occupational Safety and Health (2004) Americans work some of the longest hours in the industrialized world, and sleep less than they did in the past (National Sleep Foundation, 2005). Sleep deprivation has been indicated as a cause in 7.8% of all the Air Force's Class A mishaps (Luna, 1997). The Chernobyl, Three Mile Island, Davis-Beese, and Rancho Seco disasters all happened in the early morning, between 0200 and 0400, when powerful effects from sleep deprivation occur (Harrison & Horne, 2000). Similarly, sleep loss was cited as a causal factor that contributed to the Space Shuttle Challenger Accident (Presidential Commission on Space Challenger Accident, 1986).

Focus of the Current Study

The current study builds upon a study reported by Whitmore, Chaiken, Fischer, Harrison, and Harville (2007). Their study addressed the fact that very few studies have reported objective data regarding the effects of fatigue upon aspects of team performance, and there is a definite lack of research literature on teams in sustained operating environments (Weaver, Bowers, & Salas, 2001). Whitmore et al. trained participants on the Command, Control, and Communication Simulation, Training and Research System (C3STARS) (Reeves, Winter, Kane, Elsmore, & Bleiberg, 2001); a complex air battle management task, and then exposed them to a 36-hour period of sustained wakefulness, starting at 0300. The study quantified the effects of fatigue on ten, 3-person teams performing the complex C3 task and compared team data with individual data on a similar C3 task. The three individuals on a team each played a different role, with one

person as the Sweep, one as the Strike, and one as the Intelligence, Surveillance, and Reconnaissance (ISR). Sweep controlled fighters were used against hostile fighters; Strike controlled jammers of hostile assets and the bombers of surface to air missiles (SAMs). ISR controlled the unmanned aerial vehicles (UAVs) and the tankers, which refueled friendly, non-UAV assets. Performance data on a simple math task (Automated Neuropsychological Assessment Metrics; ANAM; Reeves, Winter, Kane, Elsmore, & Bleiberg, 2001) were also collected from the individuals and compared to the C3 task team data. Forty-minute individual and team C3 scenarios were iteratively performed throughout each experimental period, alongside traditional cognitive performance tasks. Individual performances on both the math and the individual C3 tasks significantly degraded during the early morning hours. Individual C3 data showed the well-known performance reduction resulting from sleep loss and circadian variation for both the simple and complex task levels. After sleep loss significant decrements occurred on the complex task for the C3 process measures (e.g., information gathering) and the C3 outcome measures (e.g., number of targets attacked). Surprisingly, the team data did not show the expected results. In contrast, team C3 scores on similar measures did not degrade, and in some cases showed improvements relative to baseline. These improvements were consistent with a continuing team building process.

The current report presents results of a study that involved assessing the simple cognitive and C3STARS performance of 97 subjects during a 48-hour experimental protocol without sleep and subsequently identifying, based on changes in performance, which of the participants were the most fatigue resistant and which were the most fatigue susceptible. The procedures and the findings for the analyses and parceling of the performance data are presented herein. Buccal cell specimens of all 97 participants were provided to the Allegheny-Singer Research Institute (ASRI), which has previously identified candidate genes for fatigue resistance and for fatigue susceptibility. ASRI has performed DNA extractions from the buccal cell specimens provided to them, and is conducting follow-on gene analyses to determine if the presence of the candidate genes correlates with the categorization of the subjects based on their performance.

METHOD

Participants

Ninety-seven volunteer participants, 55 males and 42 females, were recruited from the local area through e-mail and word of mouth. The mean age of the participants was 26.5 years with a standard deviation of 5.2 years. Sixty-two of the participants had either recently been in the armed services or were currently serving. The participants signed an informed consent document that had been approved by the Wright Patterson Institutional Review Board, protocol #F-BW-2006-0029-H. Participants received $500 for completing the study.

Description of Measures

Automated Neuropsychological Assessment Metrics (ANAM – Reeves, Winter, Kane, Elsmore, & Bleiburg, 2001) – ANAM is a collection of various cognitive tasks presented on a computer. For the purpose of this study, the math task, the continuous performance task, and the grammatical reasoning task from the ANAM battery were selected.

Math Task - Participants responded with left-clicks or right-clicks on a standard mouse according to whether arithmetic strings of 3 numbers (e.g., 6-4-1=?, 6-2+3=?, 1+2+1=?) had results less than 5 (left-click) or greater than 5 (right-click). No string resulted in an answer of 5. The task was self-paced, leading to more problems being presented to fast responders (there was substantial variance among the subjects on number of problems presented). Problems timed-out if not responded to in five seconds. Primary outcome measures were reaction time and accuracy (number of correct responses minus number of incorrect responses).

Continuous Performance Task (CPT) - CPT is a recognition task using single digit numbers including 0. In a stream of digit presentations (current presentation overwriting the last), participants indicated whether the current digit viewed was the same as the digit preceding the last digit (i.e., same as the one "2 back"). The pace of the task was a combination of experimenter and self-paced. If the participant did not respond, a digit would disappear after 1 second and would "time out" 1.5 seconds after stimulus onset. If the participant responded earlier than the timeout, a new digit would be presented 1 second later. Outcome measures included reaction time and accuracy.

Grammatical Reasoning Task - This task consists of 48 faceted problems, each composed of three lines of screen text: 1) a first before/after sentence (e.g., * BEFORE #); 2) a second before/after sentence (e.g., & AFTER #); and 3) a sequence of the three symbols presented in the first two sentences (e.g., & * #), where these are the only symbols used. Participants responded with a left-click, if both sentences were correct with respect to the third line or if both sentences were incorrect with respect to the third line. If one sentence was correct but the other was incorrect, a right-click was given. Time out was set at 15 seconds. This is a simple procedural task given with a fixed number of problems for each testing session. Primary outcome measures were reaction time and accuracy.

Command, Control, and Communication Simulation, Training and Research System (C3STARS) - This is conceptually faithful Air Force Command and Control simulation of a "Suppression of Enemy Air Defense" mission. There is a team mode in which three participants play in a joint-war air space and a scaled back individual-mode (i.e., with the workload scaled back to 1/3 the team condition). In this study, half the participants were assigned to the team mode and half to the individual mode. However, our final n depended on attrition in the various conditions. When a team member decided to leave the study, the teammates left behind were reassigned to the individual mode, but neither their team mode nor their individual mode data were included in the analyses.

Simulated Work (SynWin) – SynWin is a simulated work environment that requires the participant to simultaneously monitor four different activities (detailed below) on a computer screen. Cumulative work accomplished is shown center screen as the total number of points awarded on all tasks. Correct responding gains points; incorrect responding decreases points by the same amount. Time outs lose points on all tasks, except for the math task which is a "self-paced" task that never times out. The outcome metric is total net number of work points earned during a ten-minute session. The four SynWin tasks are:

Math Problem: In the upper-right quadrant, a problem requiring the addition of two, four-digit numbers is presented. The participant controlled which digits appeared in the results line by clicking a plus-button or minus-button underneath the result slot for a column (i.e., ones, tens, hundreds, and thousands). The buttons incremented or decremented the appropriate slot, and when all four slots were completed in this fashion the participant clicked a done button for that quadrant and received feedback.

Memory Problem: In the upper-left quadrant, a 6-letter memory set is given for the entire 10-minute session. After a short study time, the list disappears but subjects can re-study the list (with a score penalty) at any time by clicking its empty field. After an initial probe that has a 20 second duration, subsequent probes of the list are changed every 8 seconds (i.e., 8 seconds to respond before timing out, where time outs cost no points).

Fuel gauge: In the lower-left quadrant, a fuel gauge line moves from right to left (traversing green, yellow, and red zones). Clicking on the fuel-gauge resets the gauge to full and restarts the process. The closer one does this to the red zone, the more points are awarded. If the fuel gauge reaches zero, an auditory cue alerts the participant and points are lost for each second the fuel remains at zero.

High tone detection: In the lower-right quadrant, a big red alert button is positioned. When a high tone is detected over the head set the participant has 5 seconds to press the button to get points for successful signal detections. Lower tones are given with greater frequency and responses to these are considered errors. No points are lost for missed signals.

Psychomotor Vigilance Test (PVT) - The PVT-192 is a portable, hand-held reaction-time apparatus previously shown to be sensitive to sleep loss (Dinges et al., 1997). This task randomly and repeatedly delivers a visual stimulus to which the participant must make a push-button response within 1.5s. The inter-stimulus intervals vary from 1 to 10s. Data consist of various measurements, including false starts, reaction times to stimulus onset, and the number of lapses (failures to respond).

Profile of Mood States (POMS)-- During each testing session, subjective affect was sampled using the POMS survey (McNair, Lorr, & Droppleman, 1981), a 65-item questionnaire which, when scored according to the specified templates, measures affect

or mood on six scales: 1) tension-anxiety, 2) depression-dejection, 3) anger-hostility, 4) vigor-activity, 5) fatigue-inertia, and 6) confusion-bewilderment.

Visual Analog Scales(VAS) – Subjective mood was also measured by means of an adaptation of the VAS developed by Penetar et al. (1993). The VAS questionnaire consists of several 100-millimeter lines, each of which is labeled at one end with the words "not at all" and at the other end with the word "extremely." Centered under each line are the task adjectives (e.g., Energetic, Irritable, and Anxious). Participants indicate the point on the line that corresponds to how they feel along the specified continuum, at the time at which the task is taken. The score for each item consists of the number of millimeters from the left side of the line to the location at which the participant places his or her mark.

Activity Log -- Each participant was provided with a single sheet of paper (WFC Activity Log) on which to record their work and sleep times for three days prior to each experimental session.

DNA Sample Collection -- Buccal cells were collected with Scope® mouthwash. Participants were asked to wait at least one hour after eating or drinking before providing buccal cells. They swished 10ml of the solution 10 to 20 times before expectorating into a 50ml tube. The screw cap was replaced. One tube was collected from each participant, and labeled with their subject number. The tubes were sent by express mail at room temperature to the Center for Genomic Sciences, Allegheny Singer Research Institute (ASRI), Pittsburgh, PA. At ASRI the mouthwash specimens were centrifuged at 2,000g for 5 minutes to concentrate the cells. DNA extraction was conducted using the Puregene DNA Purification Kit specific to "DNA Purification from Buccal Cells in Mouthwash." All samples which remained after data analysis was completed were destroyed.

NEO Personality Inventory-Revised (*NEO PI-R)* – Participants completed the revised NEO Personality Inventory (Costa & McCrae, 1992) upon arrival Friday night. The inventory consisted of 240 questions and three validity questions. These questions assessed a person's personality based on the Five Factor Model: Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism.

Demographic survey – This was a short answer survey containing questions regarding the participant's age, gender, tobacco use, caffeine use, weekly sleep habits, education level, military career experience, and video game experience.

Epworth Sleepiness Scale – This survey was a self report on the likelihood that the participant would fall asleep during eight specific situations. It assessed the participant's general susceptibility to fatigue.

Fatigue Experience – The participant's past experience with fatiguing situations was investigated with this 25-item questionnaire. Questions analyzed how many times the participant had stayed awake for a full night or half night both with and without caffeine.

Also, it asked when the participant first experienced and last experienced a fatiguing situation.

WFC Sleep Behavior Questionnaire – Participants answered this 12 item questionnaire that analyzed their sleep behavior. It covered whether or not they were a shift worker, when and how long their sleep periods were, and their typical sleep amounts.

Description of Experiment

Training. Participants were required to participate in two, 4-hour training sessions prior to the experimental session. Each training session was broken into four, 1-hour segments during which the participants would rotate between the C3STARS training room and the cognitive testing room. The participants began training on Tuesday at 1800 by signing the Informed Consent Document and then being briefed on potential medical issues by the on-duty medical monitor. The participants were then split into two groups depending on the number of subjects present for that week (12 participants max.). Half of the subjects would remain in the conference room to complete study questionnaires while the other half would begin C3STARS training. Three of the members of the C3STARS group were permanently assigned to perform C3STARS as a three-member team; the others were permanently assigned to perform C3STARS individually. From this point groups would rotate each hour according to the schedule presented in Table 1.

Table-1. Training Schedule

<table>
<tr><th rowspan="2"></th><th colspan="2">Day-1</th><th colspan="2">Day-2</th></tr>
<tr><th>1-6</th><th>7-12</th><th>1-6</th><th>7-12</th></tr>
<tr><td>1800</td><td rowspan="2">C3STARS Intro</td><td>Questionnaires</td><td rowspan="2">C3STARS (2X)</td><td rowspan="2">ANAM (5X)</td></tr>
<tr><td>1830</td><td>ANAM (2X)</td></tr>
<tr><td>1900</td><td>Questionnaires</td><td rowspan="2">C3STARS Intro</td><td rowspan="2">ANAM (5X)</td><td rowspan="2">C3STARS (2X)</td></tr>
<tr><td>1930</td><td>ANAM (2X)</td></tr>
<tr><td>2000</td><td rowspan="2">C3STARS</td><td rowspan="2">ANAM (4X)</td><td rowspan="2">C3STARS (2X)</td><td>PVT & ANAM</td></tr>
<tr><td>2030</td><td>SynWin</td></tr>
<tr><td>2100</td><td rowspan="2">ANAM (4X)</td><td rowspan="2">C3STARS</td><td>PVT & ANAM</td><td rowspan="2">C3STARS (2X)</td></tr>
<tr><td>2130</td><td>SynWin</td></tr>
</table>

C3STARS training, both individual and team, was conducted over four, 1-hour training periods. It began with a general introduction of the program and its interface. Participants took part in practice scenarios that were similar to those they would conduct during the testing session. They received guidance on the most successful strategy for completing the scenarios and attaining the highest possible score. Proper ways of grouping friendly aircraft, timing on when to destroy enemy aircraft and missile sites, and proper use of surveillance aircraft in battle damage assessment were just a few of the concepts presented. The first training day consisted of the instructor teaching the participants, while the second day was more repetition in order for the participant to properly grasp the skills taught to them.

ANAM training began with a discussion on proper test taking procedures; for example sitting upright and keeping the hand and fingers on the mouse at all times. This was followed by an introduction to the three tasks they would perform during the ANAM session (math, CPT, and grammar). One cycle of all three tasks would last from 15-20 minutes. Over the course of the three hours, participants would complete at least ten cycles. Participants were given an optimal strategy for learning each task through the on-screen instructions and were reminded of the strategies by the trainer. They were told to concentrate on accuracy first, but recorded their accuracy and speed (given by the ANAM at the end of a task block). By recording the scores, the trainer was able to ensure their scores achieved an asymptote indicating that they had fully learned the task

The PVT was presented to the participants during their last hour in the cognitive testing room on the second day. Participants were instructed on the proper way to take the PVT. This included sitting forward with both hands on the PVT box, elbows on knees, and keeping the thumb on the right button. They would then complete a one minute demonstration of the actual task in order for the participant to understand what would be required during testing.

SynWin was presented during the last 30 minutes in the cognitive testing room. A total of two, 10-minute training blocks were administered. Training was initially given on the four tasks separately and then transitioned to the multi-task presentation. Participants were given an optimal strategy for learning the task through the on-screen instructions and were reminded of the strategy by the trainer. This strategy involved concentration on two of the tasks initially (i.e., math-problems and fuel-gauge monitoring), with expansion to their multi-tasking efforts as their performance on the initial pair of tasks improved.

Data collection. Participants arrived at 1800 on Friday evening to begin their 36 hours of laboratory sleep deprivation. The study schedule was broken up into nine 3-hour testing blocks. A 1-hour break was given between each testing block. During the first hour, participants ate dinner and completed the NEO-PI–R questionnaire. They were allowed to finish the questionnaire anytime before the end of the second 1-hour break. Participant sleep logs documenting their sleep for the past week were also collected during this first hour.

At 1900, the first testing block began. Participants split into the testing groups they were assigned during training. Each 3-hour testing block was divided into four, 45-minute sessions. Participants would rotate between their assigned C3STARS room and the cognitive testing room at the end of each 45-min. session. This allowed for each participant to have two C3STARS sessions and two cognitive testing sessions during each block.

Prior to starting each C3STARS testing session, participants completed a computerized questionnaire on how well they believed they would perform on the upcoming test scenario. They would then perform a 40-minute C3STARS scenario. They then completed another questionnaire that asked the participants how well they felt they

performed on the just-completed scenario. The participants would then see a screen displaying feedback from the previous scenario.

Activities in the cognitive testing room differed slightly depending on whether it was the participant's first or second session in the room during a testing block. In both sessions participants would first complete one trial of the ANAM task battery. During the first session of each block, the participant would also complete a PVT trial as well as the POMS and VAS questionnaires. During the second session, participants would complete a SynWin testing run. During each session tests began every 15 minutes. Participants were allowed to converse and play simple computer games while waiting for the next test to begin.

After the completion of the 9th testing block, subjects were released to go home. For safety purposes, participants were not allowed to drive themselves home. They were either offered a taxi drive home or had the option to be picked up by a family member or friend.

RESULTS

General Plan

The primary objective of this study was directed at "fatigue profiling" for the purpose of (eventual) correlation to genetic profiling. The second objective involved individual and team performance on the same command and control task and their respective fatigue functions. Data will be presented relevant to both objectives where it is convenient and logical to do so. However, the objectives will be covered in two separate sections that reference each other. To help with the segregation of which tests are relevant to which objective, it will generally be noted in the text and the tables which tests are being discussed (see especially "Task Choice:" section).

Primary Objective: Fatigue profiling for genetic correlation.

Task Choice:

On the basis of better training and more frequent usage in the fatigue-assessment/modeling literature, the three ANAM tasks (i.e., Math Processing, Grammatical Reasoning, and Continuous Performance Task) and the PVT (Psychomotor Vigilance Task) were given the primary objective role of supplying a behavioral profile for binning people into fatigue-susceptible or fatigue-resistant groups. Furthermore, as the PVT is a sustained attention task (e.g., vigilance) distinct from the cognitive processing tasks, this task was considered as a separate basis for classifying people.

Missing Data:

All analyses were on the 90 participants who completed the study (out of 97 who started the study). In addition tasks varied on susceptibility to missing data for various equipment/software reasons. There will be more discussion about missing data in the context of specific tasks and measures. However, the general strategy was to conduct

basic statistical analyses (e.g. general fatigue effects) for a given task by including all participants for which less than 10% of their data were missing. For our *classification analyses,* which selected a sub-sample of people who are most fatigue susceptible/resistant, simple data estimation techniques, such as using the maximum or an average on a set of *available* scores, allowed us to be more complete in our data usage (participant wise).

Task Metrics:
Given that the three ANAM and the PVT tasks were assigned to be "primary objective" tasks, we needed to decide on a best overall task metric for each to summarize participant performance on a given testing trial. For Math, participants receive more problems the faster they respond. The safest, least controversial performance measure, under these circumstances, was "right minus wrong," which weights both speed and accuracy and corrects for guessing. For Grammar, every participant received 48 problems and speeded performance was not emphasized in the instructions. Accuracy was chosen as the best metric. For CPT, problems are both participant and programmatically paced. While it is true that a variable number of CPT problems could be given with differences in response rate, this is more of a power task than is Math. CPT is also a working memory task and those sorts of tasks are usually best measured using accuracy (e.g. Kyllonen & Christal, 1990; Woltz, 1988). The decision was made to use only accuracy for CPT.

For PVT mean reciprocal response time (e.g. Dorrian, Roach, Fletcher, Dawson, 2007) and number of "lapses" (i.e., number of RTs exceeding 500 msecs, e.g. Van Dongen, 2006) are most often reported as the performance output metrics. However, our sample had substantial occurrences of "false starts" (especially in some subjects). False starts are responses occurring in the interval after a valid stimulus has been appropriately responded to but before the next valid stimulus has been presented. As both false starts and lapses significantly increased with fatigue, "lapses plus false starts" were deemed as a compromise measure for the PVT given our testing population.

As PVT is being considered as the basis of a "single-test" rule, has greater historical importance to the fatigue literature than most cognitive tests, and has (for our population) some uncertainty as to what the best metric from the task should be, the PVT is further scrutinized in Table-2. As the metrics for the other cognitive tasks are more clearly acceptable, we display results from those tasks, in the metrics we chose for them, only in figures and summary ANOVA tables.

Table-2: PVT results. Means, standard deviations, and ANOVA results for all participants.

trial	Lapses	Lapses (sd)	False starts	False starts (sd)
1	1.25	1.70	3.11	5.27
2	2.32	3.69	3.05	4.63
3	7.25	8.54	2.69	4.21
4	17.34	13.82	5.09	7.29
5	14.66	11.25	6.58	10.79
6	16.21	10.40	6.17	11.02
7	13.10	10.64	5.19	7.13
8	13.99	10.07	6.41	12.00
9	20.77	14.62	7.72	16.65
F & p	61.9	.000	5.63	.000

Table-2: PVT results (continued).

trial	1/RT	1/RT (sd)	Lapses +False starts	Lapses + false starts (sd)
1	4.13	.441	4.37	5.94
2	3.87	.520	5.37	6.65
3	3.40	.690	9.94	10.94
4	2.86	.738	22.43	18.00
5	2.97	.772	21.23	16.98
6	2.86	.682	22.37	16.48
7	3.08	.703	18.29	13.30
8	3.05	.667	20.39	15.91
9	2.65	.771	28.49	23.39
F & p	113.32	.000	49.7	.000

Note: F & p are the F-ratio and p-value from the ANOVA test for time differences.

Lapses + False Starts is not the strongest metric to consider as an index of fatigue. Response rate (reciprocal reaction time or 1/RT) had a considerably larger F for the time (fatigue) effect than lapses+false starts. If a single-rule based on PVT is desired, this metric may have to be reconsidered. However, the problem with using it in this study is that it does not consider false-start behavior (admittedly a smaller F) as part of the degradation with fatigue. Subjects with many false starts after a sustained wake are likely to show more rapid regular responding just because they have false starts (i.e., a kind of reaction time "guessing"). We will further assess our metric choice for the PVT in analyses that come later.

Basic Fatigue results:

Before describing how performance on these tasks can be used to classify people as fatigue-susceptible or resistant, general results relating to fatigue (i.e., decline in performance with later testing trials) are shown in Table-3. These are the results of

repeated measures ANOVAs for the "individual-level" tasks of the study. As expected the primary-objective tasks show *robust* time (fatigue) effects.

Table-3: Individual-level tasks

TASK	n	df /MSEs[a]	F for time effect[b]
Math Processing (Math)	88	(17,1479) / 6647, 163	40.75
Grammatical Reasoning (Gram)	86	(17,1445) / 7148, 126	56.56
Continuous Performance Task (CPT)	87	(17,1462) / 8274, 176	46.91
Psychomotor Vigilance Task (PVT)	87	(8, 688) / 6156, 124	49.70
Synthetic Work for Windows (SynWin)[2]	89	(8,704) / 6619163, 337002	19.64
C3STARS Individual (C3_I)[2]	44	(17,731)/ 90, 28	3.16

Notes:
"2" superscripts on task headings denotes a secondary-objective task.
"a" superscript on df/MSEs heading: df and MSE for the time factor and error term, respectively.
"b" superscript on "F for time effect" heading: all Fs significant at least at p<.001.

Figure 1 shows fatigue functions for the 4 primary-objective tasks. Note that PVT was tested only at 9 times (the others 18), and that PVT raw scores will show an increasing function with fatigue (rather than decreasing). Our metric choice (lapses+false starts) implies worse performance with higher scores (contra to the ANAM tasks, where high score means better performance). However, other than this one difference, all tasks display fatigue fairly similarly. One might wonder whether one could tell if the tasks did, in fact, measure fatigue across subjects *the same way* and, as a corollary, whether the fatigue decrements these tasks show fit a normative fatigue prediction model. We use an "intra-subject z scaling" technique to address the first question, and we use predictions of the Sleep Activity Fatigue Task Effectiveness, or SAFTE model (Hursh, Redmond, Johnson, Thorne, Belenky, Balkin, Storm, Miller, Eddy, 2004) to address the second question.

Figure 2 shows all of the four primary tasks on an intra-subject z scale. For example, if a participant contributes 18 scores for the Math task across the fatigue protocol, re-scaling these as "intra-subject z-scores" involves subtracting the mean of *those specific scores* from each of the specific scores and dividing the results by the standard deviation of the original specific scores. As can be seen in the figure, when the four tasks (ANAM and PVT) are rescaled this way all the functions are highly similar.

We also show in Figure 2 the SAFTE "effectiveness" predictions (re-scaled in intra-subject z units, as if SAFTE predictions were just another task given). While it is noted that the pattern of fatigue for these four tasks is qualitatively like the SAFTE prediction (i.e., decline, long-plateau, decline), the quantitative fit appears to be systematically and significantly off (i.e., just looking at the standard deviations of the tasks at the beginning and end of the protocol suggest this; no formal model fitting should be required). However, a careful reading of the source materials for SAFTE explicitly state that it is not a quantitative predictive model (at least, not in the sense ACT-r is: see Gunzelmann,

Gluck, Kershner, Van Dongen, & Dinges, 2007), but more a qualitative model which requires a further regression step for each task one wishes to quantitatively predict using SAFTE (see Hursh et. al. 2004, pg. A49). We will not be taking those further steps here, but the implication of the rescaled ANAM and PVT tasks hanging together so closely seems to be that the regression results of such a step would yield similar results for each of the four tasks shown in Figure 2.

Task Correlations:
Table-4 shows performance correlations between the tasks in Table-3, both at baseline (upper diagonal) and at fatigue endpoint (lower diagonal). Ranked performance rather than raw performance can be used to rule out any effect of outliers in the raw metrics. Outliers are possible in raw metrics but the standard deviation of ranked transformed raw scores can not vary as wildly. In fact, using raw scores, as the basis of correlations, would result in very similar, but generally *lower* correlations, than those shown in Table-4.

Correlations are generally stronger at the fatigue endpoint than at the initial baseline. This is consistent with the idea that individual differences in task performance increases with fatigue.

Table-4: Ranked task performance at baseline (upper diagonal) and at fatigue endpoint (lower diagonal) correlations

	1.	2.	3.	4.[b]	5.[2]	6.[2]
1. Math	--	.39	.40	.16[ns]	.40	.44
2. Gram	.70	--	.52	.19[ns]	.27	.41
3. CPT	.81	.73	--	.18[ns]	.46	.55
4. PVT[b]	.55	.50	.68	--	.10[ns]	.12[ns]
5. SynWin[2]	.68	.64	.70	.51	--	.64
6. C3_I[2]	.50	.46	.41	.24[ns]	.60	--

Notes:
Sample size: Math-90, Gram-90, CPT-90, SynWin-89, PVT-87, C3_I-47
Superscript b: PVT score has been reflected for positive correlation
Superscript ns: not significant, 2-tailed. All others significant at p=0.05 or less.
Superscript 2: secondary-objective task

Fatigue classification rules:
The forty-eight hour termination time for our study was determined by our desires for: 1) a fairly lengthy and challenging protocol (from the perspective of the participant) and 2) a protocol that ended when the circadian and reservoir effects of fatigue would be *maximal (i.e., early morning release)*. Because of these considerations, and the fact of a *long* plateau region in the middle of the protocol, the classification rules we considered focused on performance at the beginning and end of the protocol and excluded the data in the middle.

To make comparisons between rules more meaningful, we decided measurements on the beginning and end of the protocol should be defined the same way for any given rule considered. We considered the *maximum* of the first four protocol scores for the ANAM

tasks (and the maximum of the *first two* scores for the PVT) as the starting point, or "baseline," on which to measure changes in performance with increasing fatigue. The logic of a maximum (rather than an average) is that if participants are still learning the task (which is unexpected on our primary tasks), a later trial of the four will tend to be chosen; if the participant is fatiguing rapidly, the first trial will be chosen; and if the participant is stable, it won't matter which trial is chosen. The maximum of the first four scores is higher than what the mean of the first four scores would be but is still substantially correlated to that mean (.96, .88, .98, .95, for Math, Gram, CPT, PVT, n=90, respectively). We considered the average of the last four protocol scores (average of the last two scores for the PVT) as the fatigue performance "endpoint" to provide a good range of time and reliability on participants' fatigued performance. The last four testing trials spans the last 6 hours of the 48 hour sustained wake.

We don't claim to have explored every possible rule-classification scheme possible. However, we did explore four qualitatively different kinds and decided to apply the two most promising rules.

Percent-change rule:
This rule creates a score for each task via the formula 100**(endpoint-baseline)/baseline*, which is a metric used in the fatigue literature (e.g. Caldwell, Mu, Smith, Mishory, Caldwell, Peters, Brown, & George, 2005). In general the endpoint is expected to be less than the baseline (with the exception of PVT, which we further multiply by -1 to invert the scale); therefore, percent changes that *are more negative* show more fatigue susceptibility. When this metric is ranked across participants, participants high in rank are high in fatigue resistance. The participant percent-change was found for each of the ANAM tasks and the participant receives a rank of his/her percent change relative to other participants. Each participant's ranks are averaged over ANAM tasks and this average is itself ranked. The highest 20 scores on the averaged-rank variable are deemed the most fatigue-resistant and the lowest 20 scores are deemed the most susceptible. For PVT, simple change (endpoint-baseline) is used to rank subjects as zero is an allowable and frequent baseline score. In summary, the percent-change rule yields 20 susceptibles and 20 resistors. For PVT, there are 20 susceptibles and 19 resistors (or possibly 22 resistors, owing to ties and the inability to cut exactly at 20).

An interesting note about the percent-change rule is that, for the ANAM tasks, a percent change rule is highly similar to a rule that just uses simple change (i.e., just endpoint-baseline) and analogous ranking and aggregation procedures. A simple-change rule classifies 36 out of 40 participants the same way as percent-change rule does, with no contradicting classifications (i.e., no cases of the two rules classifying the same person differently). This may not be too surprising for this study because, for any given task, the baseline values varied much less across subjects than the values taken at the end of the protocol; hence, percent-change can approximate a simple change.

Residual-score rule:
Given our definition of fatigue baseline and fatigue endpoint, we can predict the participants' endpoint performance using their baseline performance on the same task (a

simple linear regression). We can consider the "residual" of this prediction as a kind of fatigue-resistance score (see Woltz, 1988, for an example and some discussion of this technique). This residual-score is an index of fatigued performance *not accountable* by, or controlling for, initial (non-fatigued) performance. Such residual-scores can be standardized (i.e., a z-residual) and saving such standardized residuals, as a new score, is easily accomplished using an SPSS regression option. Further, as is true of the averaging ranks in the percent-change rule, standardized residual-scores from different tasks can be averaged together without any one task dominating the variance (just as ordinary z-scores can).

A further boon for residual scores is that they readily extend their definition to other covariates. If one knows, that some demographic, such as "average weekday sleep" reported by the participant, is significantly related to fatigue endpoint performance (as we find out below), one can simply add the problematic covariate to the regression equation along with the baseline score, and have both predict the endpoint score. The residual score, from that regression, will be statistically independent of both initial performance and the covariate weekday sleep, and one can use rankings of these residuals as a "corrected" ranking for fatigue-susceptibility. For example, if someone appears to be highly fatigue susceptible in the raw data, but also reports a small amount of average weekday sleep (which is likely contributing to the fatigue susceptibility), the regression/residualization procedure adjusts that person's susceptibility ranking by removing that part of his or her score owing to weekday sleep. Note that weekday sleep is being considered as an external cause of fatigue susceptibility (i.e., a "nuisance" variable with respect to genetic causes).

We created an ANAM-based classification, in the manner of the percent-change rule, by averaging z-residuals for the 3 ANAM tasks. We ranked the average z-residual, and extracted the highest 20 ranks as fatigue-resistant and the lowest 20 ranks as fatigue-susceptible. We explore variations on how to create a residual-score in a later section, but for getting a general idea of the rule and how it relates to the percent-change rule, we consider a residual score that uses weekday sleep as an extra covariate.

We created a PVT-based residual-score model in a similar fashion; however, we only used baseline PVT scores to create the residual fatigue endpoint score, as no demographics significantly correlated to PVT fatigue endpoint for the full sample (details to be reported below).

Rule Plot comparisons:
We inspected plots and tested significance levels for susceptible vs. resistant participants identified by each rule. The expected visual differences between slopes of functions for susceptibles and resistors are evident in all rules. Figure 3 plots the percent-change rule and the residual-score rule that includes a single demographic covariate, Weekday Sleep. As can be seen the rules are very similar in the functions they produce.

Rule Overlap comparisons:
Overlap between the rules was further assessed by examining the 2 x 2 cross-tabulations between rules. For ANAM-based percent-change and residual-score rules: 18 people

were classified as susceptible by both rules, 15 people were classified as resistant by both rules, and there were no contradicting classifications between rules. However, if we consider a residual-score rule *that doesn't use additional covariates*, the overlap is even higher (19 agreements on susceptibility, 18 agreements on resistance, and no contradictions). The zero-covariate residual-score rule is hardly distinguishable from the percent-change rule.

The overlap between simple change rule and residual-score rule for the PVT was again substantial with 18/20 agreements on susceptible classifications and 17/20 on resistors. Again there were no contradicting classifications.

Finally, we considered the overlap in classifications using ANAM-based rules vs. PVT-based rules. It makes most sense to do this within rule types. The overlap between change-score type rules, ANAM vs. PVT was: 10/20 agreements on susceptibles, 12/20 on resistors, and no disagreements. The overlap between residual-score type rules: 10/20 on susceptible, 11/20 on resistors, and one case where a person is considered resistant via the PVT-rule but susceptible via the ANAM-rule. In short, the overlap between rules based on different tasks is *much less* than the overlap between rules based on the same tasks.

Fatigue correlations between tasks:
Both percent-change and residual score rules grade participants on a continuous fatigue-susceptibility score. This allows us to show correlations of "fatigue effects" between tasks for both rules. As a safeguard against outliers we rank fatigue scores (i.e., ranks of residuals or percent change) before correlating. The results are shown in Table-5. The correlations for the two different fatigue metrics are quite similar.

Table-5: Correlations between tasks based on percent-change score ranks (lower diagonal) and residual- score ranks (upper diagonal), separately.

	1.	2.	3.	4.[b]	5.[2]	6.[a,2]
1. Math		.59	.83	.58	.71	.44
2. Gram	.56	--	.70	.44	.62	.46
3. CPT	.85	.71	--	.63	.72	.52
4. PVT[b]	.59	.49	.71	--	.48	.40
5. SynWin[2]	.67	.59	.69	.51	--	.32
6. C3_I[a,2]	.44	.22[ns]	.43	.24[ns]	.34	--

Notes:
Sample size: Math-90, Gram-90, CPT-90, SynWin-89, PVT-87, C3_I-47.
Superscript b: PVT score has been reflected for positive correlation
Superscript ns: not significant, 2-tailed. All others significant at p=0.05 or less.
Superscript 2: secondary objective task

Demographics:
In preparation for considering demographics against *fatigue classification rules* we briefly investigate mean demographics for our sample of 90 protocol completers and correlations of demographics to ranked task performance.

Table-6: Mean demographic statistics.

	Mean	Std Dev	Min	Max	N
Age	26.78	5.04	19	39	90
Female	.40	.49	0	1	90
Education	2.54	.85	1	5	87
Military	.64	.48	0	1	90
Tobacco Use	.16	.36	0	1	90
Video Game Enjoyment	6.61	3.03	0	10	62
Weekday Sleep	6.90	1.05	4	9.5	90
Weekend Sleep	8.16	1.43	5	11	90
Recent Sleep	14.95	2.72	9	22	85

Notes:
Female, Military, Tobacco Use are binary variables; if the participant has the characteristic, they are scored 1; 0 otherwise.
Education: 1= High-school equivalence exam, 2= High-school, 3= Associates (Junior College) degree, 4= college degree, 5= beyond college (e.g. Masters).
Video Game Enjoyment: rating of 1 to 10 "how much do you enjoy video games?"
Weekday Sleep: hours of sleep per night "on average" during the weekdays
Weekend Sleep: hours of sleep per night "on average" on weekends
Recent Sleep: hours of sleep reported for Weds and Thurs prior to the study (Friday).

Sample sizes on demographics vary because of missing data. However, the low n on Video Game Enjoyment reflects asking for this demographic after we perceived a recurring trend for participants to say they liked some of the tasks (e.g. C3STARS and SynWin) because they liked video games. Our n on this demographic reflects only participants who filled out the questions assessing video game perceptions, which was added to our demographics questionnaire after the first 24 participants had completed the study.

For the variables Weekday and Weekend Sleep, it makes sense to assess their mean difference, which is in fact quite significant: $t(89)= -7.25$, $p<.0001$. These variables are also uncorrelated ($r(90)=.16$). The fact that our participants report more weekend sleep than weekday sleep suggests they use the weekend for recovery sleep and do not get their preferred amount of sleep during the week.

The variable Recent Sleep was chosen to extend 2 nights before the study (instead of 3) owing to missing data on the forms that recorded sleep behavior. More data was missing for Tuesday night (when we first gave them the form). Therefore, Recent Sleep is hours of sleep reported Wednesday and Thursday night, plus any naps they report on Wednesday, Thursday, or Friday day. Naps were infrequently reported. If we assume Recent Sleep is a sample of 2 nights of "weekday sleep," another informative analysis is to divide Recent Sleep by 2 and compare that estimate of weekday sleep to participants' reports of "on average" Weekday Sleep. In fact, their recent sleep estimate was significantly greater than their reported "on average" estimate ($t(84)=-3.25$; $p<.002$, comparing the means 6.91 to 7.48), and these two variables were uncorrelated

(r(85)=.14). This suggests participants slept (a little) more than usual just prior to the study.

Table-7: Significant correlations between demographics and task-performance ranks at baseline and endpoint of the fatigue protocol.

Demographic	Baseline	Endpoint
Age	C3_I: -.44; SynWin: -.27	None
Female	C3_I: -.36	None
Educ	C3_I: -.36	None
Military	None	None
Tobacco	Gram: -.21	None
Video Game Fun	C3_I: .61; SynWin: .40; CPT: .26	None
Weekday Sleep	CPT: .25	CPT: .31; Math: .22; Gram: .21
Weekend Sleep	None	None
Recent Sleep	PVT: .27	None

Notes:
Alpha ≤0.05 (2-tailed) was the criterion for a significant correlation.
Sample size: Math-90, Gram-90, CPT-90, SynWin-89, PVT-87, C3_I-47.
PVT score is reflected for positive correlation.

Table-7 lists Task/Demographic correlations, highlighting significant correlations. Recall that in this section we are mainly focused on Math, Gram, CPT, and PVT. A two-tailed .05 significance level is adopted. Correlations are observed at both baseline and endpoint of fatigue protocol. Some demographics correlate to baseline performance but not to endpoint performance. These demographics are the ones typically found in the non-fatigue performance literature (e.g. gender, age). However, the Weekday Sleep demographic has the reverse trend and correlates *more to endpoint*. As this demographic could reflect amount of *chronic* sleep deficit, perhaps its effects should be more apparent after a sustained wake than before it, so the reversal for the Weekday Sleep demographic, relative to the other demographics, makes sense. Recent Sleep (participant reports of sleep just prior to the study) correlated to PVT performance at the beginning but not at the end (where r(82) = -.06 is the correlation at the end).

PVT *endpoint* and baseline appear to correlate to Weekday Sleep, like the ANAM tasks, but not up to our stated acceptance levels (r(90)= .15, p<.161, for baseline; r(87)=.18, p<.096,). Given that we expected PVT to be *more sensitive* to fatigue than the ANAM tasks (based on the literature), and that we found the ANAM tasks showing greater sensitivity to fatigue via their more robust correlation to weekday sleep in this study, we considered the PVT more closely.

We looked at PVT and ANAM task reliabilities at fatigue endpoint. Using trials 8 to 9 correlations to gauge the reliability of the PVT endpoint score, and the average of the trials 15 and 17 correlated to the average of trials 16 and 18 to gauge the reliability of an ANAM-type score, we note split-half correlations .41, .79, .76, .88, for *ranked* scores on PVT, Math, Gram, CPT, respectively (for raw scores; .55, .77, .80, .86, respectively). Hence, *lower reliability* of PVT relative to ANAM is a factor. We also looked at the

consequences of our metric choice for PVT. In particular, we examined whether PVT endpoint performance would be more sensitive to Weekday Sleep had we used another PVT metric, such as mean reciprocal response time, lapses alone, or false starts alone. The answer is no: the four possible metrics (including our metric choice) correlate from .16 to .19 (all not significant).

As another kind of task which would help us assess the quality of our PVT metric choice, consider the average of the ranks of percent-change for the 3 ANAM tasks (i.e., that variable from which we draw the top 20 resistors and the bottom 20 susceptibles for the ANAM-based percent-change rule). This variable can be considered a kind of "construct-validity" variable that any fatigue measure, such as the PVT can "predict." Now consider the rank of the simple change scores (i.e., endpoint-baseline) for the four possible PVT metrics and how they correlate with this construct validity variable. The rank of this change-score for our chosen metric (lapses + false starts) correlates the most ($r(87)=.66$), and mean reciprocal response time correlates next highest ($r(87)=.61$).

Correlations of classification rules to demographics:
Rules classifying people can relate *significantly* to demographics and this can be tested using two statistically equivalent methods. The first is a t-test for a group difference between resistors and susceptibles for a given rule on the demographic; the second is coding resistor as 1 and susceptible as 0 for a given rule and correlating this new binary variable to the demographic. We'll use correlations but also give mean differences, with significant findings, for a better appreciation of the size of effect. Our stated level of significance is $p<.05$, two tailed, but we will also comment on "near misses."

For the percent-change ANAM-based rule: Greater Weekday Sleep goes with being fatigue resistant: $r(40) = .36$, $p<.02$, means 6.62 for susceptibles vs. 7.38 for resistors. We can also look at the correlation of Weekday Sleep to average ANAM percent-change rank over all 90 participants, $r(90)=.25$; $p<.015$, or consider a more extreme set of subjects--i.e., the top 11 and bottom 11 participants, $r(22)=.47$, $p<.028$, means 6.64 vs.7.55. Hence, the relationship between Weekday Sleep and this rule is robust in the sense of being present on the entire sample and on two different definitions for our extreme groups.

No significant effects were found at our stated significant levels for the PVT-based simple-change classification rule; however, the two "near misses" (i.e., 2 highest relationships observed) were: Weekday Sleep, $r(39)=.26$, $p<.11$, and Education $r(38)=.22$ with both correlations in the expected direction. If we counted these results significant, it would mean people with more weekday sleep and higher education do better on PVT.

For the ANAM-based residual-score/1-covariate rule (using Weekday Sleep as the covariate), we note that by statistical correction, the correlation of this rule to average weekday sleep reported should be near zero and it was (average hours, 6.90 vs. 6.72 for susceptible vs. resistant). This rule did not correlate to the other demographics at the stated level of significance; however, the correlation to Recent Sleep, which was the largest non-significant correlation, was suggestive at $r(40)=.25$, $p<.11$, two-tailed. That

being the case we re-defined the residual score to include recent sleep as an additional covariate. To keep our sample size for this newer rule the same (i.e., n=90) as the residual-score/1-covariate rule, participants with missing data on Recent Sleep (i.e., 5 participants) were assumed to have the mean value of Recent Sleep. With this adjustment the means on weekday sleep for susceptibles vs. resistant extreme groups are -- weekday sleep: 6.80 vs. 6.98, $t(38) = -.60$, ns; recent sleep: 14.87 vs. 14.62, $t(38)=.26$, ns.

In may be instructive to trace exactly what is happening with the residual-score rule as various options for it are implemented. Starting from the case of no covariates (i.e., just using the baseline to create the residual-score for fatigue resistance), the overlap between the residual-score and percent-change rule is very high, and therefore, this zero-covariate residual-score rule also relates to the Weekday Sleep demographic as did the percent-change rule (i.e., $r(40)=.32$, $p<.023$). When Weekday Sleep is added as a covariate in the residual-score creation and the participants are reselected on the basis of the changed rankings, Weekday Sleep no longer relates to rule classification ($r(40)= -.08$, $p<.60$, note that resistant people now report slightly less sleep). As reported before, the percent-change and the one-covariate residual-score rule overlap 18/20 on susceptibility and 15/20 on resistance. Therefore, to go from the percent-change selections to the revised residual-score/1-covariate selections, 7 people were removed from the percent-change rule's classifications and 7 new people, who are unique to the 1-covariate residual-score rule were substituted for them.

Next the 2-covariate residual-score rule is considered and Recent Sleep is added as a covariate (along with Weekday Sleep) to create a new fatigue resistance score. The 2-covariate rule overlaps with the 1-covariate rule 19/20 on susceptibility classifications and 17/20 on the resistance classifications, so a change of selection occurs by removing 4 participants from the old 1-covariate rule and substituting 4 new participants, unique to the 2-covariate rule. For the two-covariate rule, correlations to Weekday Sleep go up a little, $r(40)=.10$, $p<.550$, while the correlation to Recent Sleep goes down, $r(40)=-.04$, $p<.80$. It is interesting to note that the two-covariate residual-score rule overlaps slightly *more* with the percent-change than the one-covariate rule does (18/20 susceptible overlap, 16/20 resistant, for the 2-covariate overlap with the percent-change rule). In other words, adding covariates doesn't always lead to decreasing overlap with a specific reference rule.

For PVT-based residual-score rules using the two sleep covariates, there were two near misses to our stated levels of significance for Tobacco, $r(40)=.28$, $p<.08$ and Video Game Enjoyment $r(27) = -.31$, $p<.11$. The Tobacco Use correlation (actually a near miss) is hard to interpret given participants were not allowed tobacco during the study, as specified by the Informed Consent Document, and given Tobacco Use hurts Grammatical reasoning performance (contra PVT).

Secondary Objective: Fatigue and teams; fatigue and tasks.

Task Choice and Descriptions:

For this section we are interested in the Command and Control task, C3STARS. Consideration of team and individual C3STARS performance conditions can be used to assess whether fatigue functions are similar for teams and individuals. One should recall that our manipulation of "teaming" is quite subtle. Teams and individuals play in a similar laboratory room and do exactly the same task-work but teams fight in a common war (and in a common geography) and can *elect to* coordinate with each other.

We are also interested in the SynWin task as a companion task to C3STARS in individual mode. In the case of C3STARS we were unable to train all participants reliably to asymptote (although some participants achieve ceiling performance on the individual task, even before the protocol starts). In the case of SynWin, we intentionally train only slightly to make this task more C3STARS-like, that is, *not fully learned* at the start of the protocol. Even so, SynWin's components were well understood at the beginning of the protocol, and it is (arguably) just the management of ones cognitive resources in doing all 4 tasks simultaneously that becomes the learning objective during the protocol. This is directly analogous to our participants being better-trained on C3STARS mechanics than the same mechanics while multi-tasking during scenario play. However, with C3STARS one can expect more complex learning than just how to manage time (e.g. perceptual learning, such as how far an asset can be from a jammer while still being safe), so we had no strong a priori hypotheses that C3STARS and SynWin would actually resemble each other under fatigue.

C3STARS Missing data compared to other tasks:
C3STARS individual condition had missing data comparable in frequency to the ANAM and for the same reasons (e.g. rare equipment failures). Missing data for C3STARS was primarily a problem for the team condition. There were several causes of missing data in the team condition. A team-specific issue involved subject attrition. For the seven subjects who elected not to finish the study, five of these were from the team condition (2 from the same team). This meant that of the 16 teams that started the protocol, 4 teams could not be analyzed, because the team broke up before the protocol ended. When a team loses a teammate, the remaining teammates convert to the individual-mode C3STARS, and their previous team data and subsequent individual data for C3STARS are discarded.

Equipment and software problems in the C3STARS team-condition were larger than for other tasks. Only 5 of the 12 protocol-completing C3STARS teams have complete data (i.e., all 18 trials). The majority of missing data arises from the distributed-simulation (i.e., team) version of C3STARS being less forgiving to proctor mistakes. For instance, not saving the data or prepping the next run correctly can cause the simulation to crash or act buggy in the next administration, requiring a restart. If a restart occurred substantially into a scenario, the trial data became missing, as partial plays could not be used. In addition a smaller source of missing data in the team condition should be noted. In 3 of the 216 team-trial blocks (but in none of the 792 individual-trial blocks), a friendly Jammer, programmed to be invulnerable to normal (as opposed to No-Fly-Zone) SAMS, was shot down by a normal SAM. As jammers are valuable assets and losing them

usually involves losing associated assets, we decided to treat these 3 team blocks (from 3 different teams) as missing data also.

To assess the impact of missing data in our team vs. individual repeated-measure contexts, we will use multiple methods and sample sizes. In particular, we will compare analyses where missing data are filled in by data imputation (team n=12) to analyses performed on only full-data teams (team n=5). We will also perform more focused analyses, involving fewer trials (e.g. the most fatigued trials), to obtain a larger number of full-data teams over the analyses. We think comparisons between team and individual C3STARS replicate nicely over imputed and full-data subsets, but the only way to show this is to provide the comparisons.

Task Metrics:
For C3STARS the metric was a single mission outcome score made up of adding and subtracting points awarded for "good and bad" scenario events in the course of play. Shooting an enemy was a good event, though only enemy aircraft died with certainty when shot. Enemy SAMs needed a successful battle-damage assessment to confirm a SAM was successfully shot. Therefore, participants received 1 point for each distinct enemy shot and 1 extra point for each SAM successfully assessed (which implies the SAM was shot, and shot the required number of times). Participants lost 1 point for every friendly asset that died during the session. While friendly losses are not directly weighted by the value or importance of the asset lost, a lost asset *will be weighted by the simulation outcome* according to the severity of that particular loss. As an example, consider that losing a bomber *before* its bombs have been used is a worse mistake (in terms of final-scenario score attainable) than losing it after the bombs have been used. As another example, consider that losing a jammer often means that the rest of the strike package that accompanies it are killed by enemy SAMs.

For SynWin the metric chosen is provided by the task creator and is a total score based on the amount of work done during the trial period. Each time one of the four SynWin tasks is done correctly positive points are added to the total; each time an incorrect response is given, points are subtracted. Three of the four tasks have time outs leading to point subtraction on ignoring them; however, the fourth and hardest task (multi-column addition) is self-paced. Recall that the total work score is provided to the participant continuously as they do the task.

Fatigue Functions: Teams vs. Individuals on C3STARS.
We consider fatigue in the context of being on a C3STARS team vs. being in the individual-play condition. Figure 4 gives a big-picture overview of raw data from both conditions. However, the team condition is displayed via two options: 1) only teams that had full data (i.e., all 18 trials) as the pink function, and 2) data from all 12 teams, where 7 teams have imputed data values on their trials which had missing data. The method of imputation was to average the extant value of the trial just preceding and just succeeding the missing trial, and use that average as the estimate for the missing trial value. Finally, the individual condition (yellow function) has only those participants having complete data (all 18 trials). As missing data due to equipment/software is much less of a problem

with the individual C3STARS condition (i.e., 44 out of 47 possible have full-data), we decided imputation for the individual condition would not be beneficial.

The individual condition is substantially lower than the team condition, which reflects differences between the tasks and not the subjects; i.e., the "war" in the team condition is 3 times larger than the war in the individual condition. The take home message from the figure is that team-performance is larger than individual performance throughout most of the protocol, but by the very end of the protocol teams output as much work (measured in mission outcome) as individuals do. This certainly seems to suggest that teams are crashing harder from fatigue (proportionally speaking) than individuals do.

Table-8 shows the means relevant to Figure 4. The first set of columns, "… means for available data …" shows a different view on our team missing data problem. In particular, if we just consider the n-column for available data, the missing data are tolerably (i.e., uniformly) distributed across the protocol and no trial number is missing more than 2 team's worth of data.

Table-8: Means, standard deviations, and ns for data comparing teams to individuals shown in Figure 4.

	12-teams: means for available data (n displayed for each trial)			Teams with imputed missing data (n=12)		Teams having full-data (n=5)		Individuals having full-data (n=44)	
Trial number	mean	sd	n	mean	sd	mean	sd	mean	sd
1	19	13.15	12	19	13.15	14.60	15.53	7.05	7.73
2	26.17	11.17	12	26.17	11.17	27.20	7.92	8.09	7.29
3	24	12.66	11	23.33	12.29	21.80	5.72	6.7	8.22
4	25.55	12.62	11	24.75	12.34	25	10.49	8.2	7.49
5	20.58	13.49	12	20.58	13.49	17	13.4	8.5	6.75
6	22.42	13.94	12	22.42	13.94	14	7.45	8	7.97
7	22.9	15.66	10	23.5	15.05	15.20	12.48	6.98	7.85
8	22.67	10.24	12	22.67	10.24	18.60	9.53	7.77	7.5
9	20.75	12.61	12	20.75	12.61	20.4	9.84	8.55	7.01
10	24.67	12.23	12	24.67	12.23	21	13.10	7.61	9.17
11	25	12.18	12	25	12.18	21.40	10.69	7.61	8.77
12	19.5	12.03	12	19.5	12.03	19.6	8.73	9.2	7.44
13	20	16.85	10	21.04	15.54	18.2	13.75	9.64	7.69
14	29.27	14.48	11	27.79	14.73	27.6	15.18	9.57	7.86
15	23.92	13.75	12	23.92	13.75	17.8	10.28	8.02	7.62
16	15.91	16.24	11	16.83	15.82	13.6	12.54	5.57	8.24
17	13.20	10.17	10	13.83	10.96	12.8	8. 61	5.55	8.83
18	4.67	9.83	12	4.67	9.83	2.00	11.40	4.18	8.07

To explore the fatigue functions and possible differences between teams and individuals with clearer interpretations, we consider repeated-measure ANOVAs. One manipulation that helps with interpretation and data exploration is to break up the 18 time levels into 3 different epochs (i.e., trials 1-6, 7-12, and 13-18). Hence, the analysis has two within-subject factors, namely epochs (i.e., early, middle, and late in the protocol) and time (i.e.,

the 6 trials points within each epoch); and one between-subjects factor (i.e., whether the "participant" is a team or an individual).

Table-9 shows two separate ANOVAs, one using imputed data (team n=12) and the other using the full-data subset (team n=5). Both analyses compare their respective team condition to the same individual condition (individual n=44).

Table-9: Repeated-measures ANOVA results comparing teams and individuals raw-performance under fatigue (time and epoch effects).

EFFECT	F(i)	F(f)	df(i)	df(f)	p(i)	p(f)
team	34.83	12.02	1, 54	1, 47	.000	.000
epoch	11.05	4.98	2, 108	2, 94	.000	.000
team x epoch	5.71	2.50	2, 108	2, 94	.004	.088
time	10.88	8.51	5, 270	5, 235	.000	.000
team x time	6.70	6.07	5, 270	5, 235	.000	.000
epoch x time	10.73	6.28	10, 540	10, 470	.000	.000
team x epoch x time	3.95	2.20	10, 540	10, 470	.000	.017

Notes:
Column headings ending with "(i)" are for the imputed-data (team n=12) analysis.
Column headings ending with "(f)" are for the full-data sub-sample (team n=5) analysis.

With the exception of the team x epoch interaction (which just approaches significance), effects present in the imputed analysis replicate in the full-data sub-sample. We briefly describe the meaning and importance of some selected effects. The team effect as a main effect is uninteresting; however, its interaction with other effects could be relevant to whether teams fatigue as individuals fatigue. The epoch effect says that, independent of the team effect, performance varies by epoch (i.e., fatigue effects). The team x epoch interaction (significant only in the imputed sample) says that the difference in performance between teams and individuals varies by epoch. Not surprisingly (see Figure 4) the difference between teams and individuals is least in the last epoch. However, we can't use this significant F to support the statement "teams fatigue more than individuals," because a theory, which posits team members doing the same amount of work, and fatiguing as an individual player does, *also* predicts significant *team x epoch* interactions. Figure 5 provides a schematic for this theory.

In summary the analysis, to this point, is useful only insofar as: 1) It shows we have sufficient power to detect main effects and interactions in the raw data. 2) The main effect of epoch is significant giving statistical backing for fatigue occurring at the end of the protocol for C3STARS. 3) With this raw data, we can also support the idea that fatigue happens *primarily* in the last epoch. We can do this by manipulating the focus of analysis; in particular, if we remove the last epoch, and redo the analyses with just the earlier epochs, neither epoch main effects nor team x epoch interactions are significant (e.g. Fs<1 for the imputed-data analyses).

As the analysis stands, we haven't shown that the team endpoint (being close to the individual endpoint) indicates teams are more vulnerable to fatigue, because, in the raw data, it is the size of the team x epoch interaction, and not just its existence that would indicate this. Therefore, we leave the raw data, and do a "proportional change" assessment of teams and individuals under fatigue to assess this ambiguity. This assessment should have the best chance of working at the last trial. If the analysis is not significant there, where the convergence of the team and individual functions is most apparent, it shouldn't become more significant by aggregating earlier fatigue points into the comparison (teams vs. individuals), as these earlier points should lie closer to where the team function ought to have ended up, given Figure 5 is the true state of affairs (that we have measured with substantial noise).

There is more than one method to put team and individual performance on a "proportional-change" scale of fatigue-impact. One way to do this is to compute the percent-change of the final performance (in the current case, the last point) from a reasonable baseline. For the ANAM tasks, we used the maximum of the first 4 trials, as a baseline (as the maximum appeared to be somewhere early in the protocol as expected). However, for C3STARS, it is not clear what stretch of performance should be used to define a baseline. We, therefore, chose our "baseline" to be the maximum, or the mean, over the first 15 trials, as only for the last 3 trials of both C3STARS conditions seem clearly to be degraded from fatigue. For each participant "unit," be they individual or team, we convert their last performance point to a percentage-change from this baseline. One then tests whether this percent-change score differs between teams and individuals, considered as two different groups.

A second way to create a "proportional" fatigue-impact scale is to compute a residual score for all units. To compute the residual score, one would predict last performance by the maximum, or mean score, on the first 15 points and then do the team/individual comparisons on the residual scores. Note that whether a unit is a team or an individual is not *directly* co-varied out (if it was, then *the mean difference* between teams and individuals on the computed residual-score would be zero). However, a unit's final score *is being predicted by* its baseline score, so this takes into account the fact that if a unit is a team, it tends (on average) to have a higher score than an individual. Residual-scores also replicate the percent-change logic. To put it as if these scores could actually speak: the residual-score asks: "how deviant is this unit's final performance relative to my expectation of its final performance based on taking into account its initial performance"; whereas, the percent-change score says: "how big is the change from where this unit started, scaled by (divided by) where it started".

Finally, a third way to do the test is to convert each unit's 18th fatigue point to the intra-z-score scale (as defined earlier) and do the team/individual comparison on that metric for that point. This assesses the distance a unit's most fatigued score is from the unit's mean score, scaled by that unit's performance variance over the protocol. It's interesting to note that this kind of score doesn't have the same kind of "baseline" component or baselining step that the other two scores have.

For methods that have the option of using a maximum or a mean, we report the option that works best and comment on the other, less successful, option. As all teams and all individuals have trial 18 (the last point), and as we just use available data to compute baselines, means, and standard deviations, our team-n is 12 and our individual-n is 47 for these analyses. The results are shown in Table-10.

Table-10: Assessments of Team vs. Individual Fatigue impacts via meaningful change-score metrics (see text for definitions)

Metric/unit-type	mean	sd	n	t/p
Percent-change (from max): Individual	-79.0	88.1	47	t(57) = .42 p<.679
Percent-change (from max): Team	-90.0	26.4	12	
Residual score (by mean): Individual	.149	.918	47	t(57) = 2.38 p<.021
Residual score (by mean): Team	-.585	1.092	12	
Intra-unit z-score: Individual	-.544	1.013	47	t(57) = 3.87 p<.000
Intra-unit z-score: Team	-1.739	.655	12	

Notes.
P-values are for two-tailed tests.

For percent-change, using a maximum baseline, rather than a mean baseline gave better results, even though neither option was significant. Using the mean option actually leads to mean differences in the wrong direction (i.e., individuals fatigue more than teams). One problem with using percent-change scores is a division by a baseline. If the baseline is near zero for some participants (which is allowable in C3STARS, as are negative scores) variances on the percent-change metric may blow up. Something like this may be happening for the individual group, under the mean option, where the standard deviation on percent change was 308 percent.

In contrast to percent-change, residualizing by the mean of earlier scores works better than residualizing by the maximum of the earlier scores, though the latter was in the correct direction ($t(57)=1.60$, $p<.12$). That the mean works better than the maximum for residual scores follows from the mean-score baseline predicting the last score better than the maximum-score baseline ($R(57)=.46$ vs. $R(57)=.29$, respectively). As we consider the baseline score "nuisance" variability that we want the last score "corrected for," a smaller R for the maximum baseline measure (compared to mean) means poorer correction for that nuisance variability (compared to the mean), and with poorer correction for a confound, one generally finds less significant effects.
Finally, the most significant metric, supporting the statement "teams were more adversely affected by fatigue than individuals," is the intra-subject z metric. Therefore, we decided to re-plot the means and re-do some analyses under this metric. The plots are shown in Figure 6. Rather than do an 18-trial analysis, we focus at the end (last epoch and the last 6 trials) to see whether the slopes significantly differ. In the raw data they certainly

would differ. In this newer metric, however, and under an assumption that a model like the one in Figure 5 *is true*, there would be no slope difference (or main effects) observed.[1] For this analysis, we only use full-data teams within the last 6 trials (n=7) and and full-data individuals (n=44). Both a time and team x time interaction is significant ($F(5,245)=10.10$, $p<.000$; and $F(5,245)=2.26$, $p<.05$, respectively). This analysis (on just 7 teams) agrees with the analysis of just the last performance point (which uses all 12 teams). The two analyses, together, provide statistical significance to back up the idea that team performance degraded more in the same time period from fatigue than individual performance did (i.e., the interaction that seems compelling in the raw data really is larger than what the model portrayed in Figure 5 could predict).

Basic Fatigue Results: Types of Tasks

Here we concentrate on the individual-mode C3STARS task (hereafter, C3_I) and how its performance changes with fatigue. The general question addressed is whether C3STARS is a member of a *family of similar tasks* that seem more resistant to fatigue, or at least different with respect to fatigue, when compared to the family of tasks more typical of the fatigue literature, e.g. cognitive tasks (ANAM) or sustained attention tasks (PVT).

Correlations:

Table-4 shows that C3_I and SynWin correlate to other cognitive tasks, but in particular C3_I and SynWin are most highly correlated to each other, at least when participants are rested. This high correlation suggests more similarity between these two tasks than to the others. The correlations shown in Table-7 also suggest some similarity. C3STARS and SynWin both correlate to the age demographic (while no other tasks do) and both have a relatively high relation to Video Game Enjoyment (shared by Grammatical Reasoning, though its correlation to this demographic is least).

Fatigue Functions: Types of tasks

For this analysis we use the even trials of C3_I (which are closest to the nine SynWin administrations) and compare the fatigue functions in the intra-z score metric (for C3_I all 18 points are used in z-scoring of the even points, although the results reported below replicate, if we just use the 9 even points). Results are shown in Figure 7. Here we can see more direct evidence of a C3STARS/SynWin "family resemblance." By contrast PVT (in the intra-z metric) is also plotted to show the difference between C3_I and SynWin from primary-objective tasks. In particular, C3STARS and SynWin resemble each other much more than either resembles the PVT (which we showed in Figure 2 to resemble the other ANAM tasks in this metric scale).

In exploratory analyses in the intra-z metric, we compare C3_I to SynWin throughout the

[1] The easiest way to see this is to apply the intra-subject z scaling to two hypothetical participants having very different slopes and main effects across 3 levels: e.g. 30, 20, 10 for one "team-like" participant and 10, 6.67, 3.33 for another "individual-like" participant. The difference between the participants at each level will be 20, 13.33, and 6.67 in the raw data. In the intra-subject z scale metric the difference between the 3 participants will be zero (within rounding error) at each of the 3 levels.

protocol (9 levels of time). There is a significant time effect ($F(8,368) = 9.27$, $p<.0001$) and a modestly significant time x task interaction ($F(8,368)=2.53$, $p<.011$). The latter interaction suggests something like the interaction that appeared between team and individual C3STARS conditions, i.e., SynWin appears to be changing from fatigue at a faster rate at the end of the protocol than C3_I is.

We can further explore the time x task interaction, by using the procedures we used before to explore similar issues concerning team vs. individual C3STARS fatigue-function differences. First we divide the 9 times into 3 epochs and note that with a repeated measures analysis (within-subject factors: epoch, time within epoch, and task) the following significant interactions bear on the question "does performance on C3_I drop off at a different rate than performance on SynWin under the same fatigue?": epoch x time ($F(4,184)=17.55$, $p<.0001$) and task x epoch x time ($F(4,184)=3.54$, $p<.008$). The two factor interaction supports the suggestion from Figure 7 that fatigue is mainly occurring in the last epoch (i.e., this is the best place to look for task differences between C3_I and SynWin with regard to fatigue), and the three factor interaction supports the suggestion that performance drop-off is happening more slowly in C3STARS. Next we do another analysis that leaves off the last epoch and note that the previous interactions are no longer significant (maximum $F(2,92)=1.82$, $p<.168$). This gives us *some* justification for re-focusing the analysis to just the last epoch (i.e., the last 3 times). In this more focused analysis, analogous tests are significant, namely a main effect of time ($F(2,92)=28.45$, $p<.0001$) and a time x task interaction ($F(2,92)=4.99$, $p<.009$), and they bear analogously to the quoted question above. Here the time effect is a fatigue effect across tasks and the interaction is a differential impact of fatigue across tasks. At least we speculate this interpretation is valid. It depends on the validity of the intra-subject z-scale as being the most appropriate metric to be assessing this question.

DISCUSSION

Primary Objective:

Our main results here are the assessments of procedures for classifying who, in a population sample, are the most fatigue-resistant and most fatigue-susceptible, where we want the classification to be more likely owing to genetic causes (rather than external causes). We discovered that task-performance profiles that indicate a person as susceptible or resistant (e.g. as shown in Figure 3) can be multiply-caused. There could be genetic factors determining ones task-performance profile, but life choices, such as the amount of weekday sleep one elects to have, can also contribute to the profiles. We made the assumption that, weekday sleep, the most influential of our demographics with respect to our fatigue classifications, is independent of the genetic factors that are targeted for follow-on correlational analysis. We found that using a residual-score (rather than percent-change score) could be used to create fatigue classifications that were more adaptive to our goals of excluding such external (non-genetic) causes for our classifications. The residual-score approach is arguably better than manually matching people (e.g. in susceptible and resistant groups) on one or more covariates and then performing the genetic analysis only on the sub-sample that could be matched. This manual procedure can reduce the n for analysis substantially.

With regard to the residual-score rule a few other observations are useful. First, the residual-score, created by just using an initial performance covariate, showed high overlap to the percent-change rule with regard to subject selections. This suggests that had a residual-score rule been used in past literature, which uses percent-change to assess fatigue impact (e.g. Caldwell, et. al. 2005), the results reflected by that literature would not change very much. Residual-scores also are not afflicted by certain boundary conditions that show up with percent-change scores (e.g., zero or negative numbers as allowable baselines). As a caveat, one does have to be mindful that severe outliers can skew the residual-score more (relative to a percent-change option), because such might shift the regression line of best prediction substantially, and this would seem to have an effect of changing the residual-scores for the rest of the sample (although, we are less sure of what precisely the effect would be on residual-score *rankings*). We did not see any outliers of the overly distortive kind that had to be thrown out. We also think that the initial-performance-only residual-rule would not have "tracked" the choices of the percent-change rule so well, if there were super-distortive outliers in the sample. Extreme outliers, under a percent-change rule's perspective, would just rank such outliers at the top (or bottom) of the distribution but would not affect the ranks of people between them.

Another caveat for the residual-score rule is precisely when to residualize by a demographic. The answer may be more uncertain when the demographics involved are not obvious confounds. Suppose a demographic like Video Game Enjoyment *had been* significantly related to fatigue classification on the PVT. Should that demographic be co-varied out of a PVT-based residual score? We don't know, because we don't know whether that demographic should be considered a confound relative to some hypothetical genetic trait or a common expression of such a trait. What causes video game enjoyment? Perhaps the need for stimulation does. What causes poor PVT performance? Fatigue certainly does, but perhaps the need for stimulation or to be engaged in a task does also. How genetic is the need for stimulation? We don't know.

Secondary Objective:

The secondary objective is an exploratory assessment of moderators for the amount of fatigue a person *shows* when task context varies. Fatigue, according to normative models (e.g. SAFTE, Hursh, et. al. 2004), is not a property of the context one measures fatigue in, but is a consequence of amount of time without sleep and the time of day. However, the impact of fatigue seems to differ according to context. In particular, for this study we have to consider: 1) Why do teams seem to degrade more in their performance under fatigue than individuals, at least at the end? 2) Why do tasks like C3STARS and SynWin exhibit fatigue differently than primary-objective tasks? We will be looking at candidate answers to these two questions and speculate on what further kinds of studies might bear on them.

The apparent finding of teams getting hit harder by fatigue was counter expectations deriving from Whitmore, Chaiken, Harrison, Harville (2007). To review the main

empirical results of that earlier study of interest here: 1) Teams that played C3STARS continued to learn during the fatigue protocol; in fact, the team condition's most-fatigued point (after about 30 hours of sustained wake) had performance near a second baseline point taken after protocol recovery (i.e., after the study ended and about a 16-hour recovery sleep). 2) The individual-mode C3STARS condition showed a significant mid-protocol drop in performance and then a substantial rebound (to near baseline performance) at the end of the protocol (i.e., after 34 hours of sustained wake). This recovery was not a short-lived "going home" effect but lasted through 2 testing points (about 2 hours). 3) Teams did not do as well as individuals, when compared on percent of maximum score (percent of perfect mission outcome) they attained. The team condition was a harder condition in this respect. And finally 4) Learning was largely absent from the individual condition as baseline (pre-fatigue performance) was equivalent to recovery performance (i.e., performance after the recovery sleep). So for the individual- C3STARS condition, it can be said that asymptotic performance was established prior to the fatigue protocol.[2]

There were also many procedural differences in the earlier study. We list some of the more important differences (without knowing the order of importance): 1) There were fewer testing trials for all tests (ANAM and C3STARS) in Whitmore, et al. (2007), leading to a much higher down-time/work-time ratio. C3STARS-type tasks were also given much less frequently in Whitmore et al. (2007), which had 4 and 5 total fatigue-protocol trials, for team and individual conditions, respectively. 2) Whitmore et al. (2007) used military participants *from the job domain* of C3STARS. 3) A much longer training period was given in Whitmore et al. (2007) (3 complete days vs. 1 day over 2 evenings). 4) A different team architecture was employed in Whitmore et al. (2007), namely a functional architecture, in which team players controlled only one type of asset, and teammates had to coordinate usage when the scenario frequently required multiple types of assets to do the job (i.e., one team-mate controlled jammers and bombers, another the air-to-air fighters, and the third, the ISR assets). 4) A shorter fatigue-protocol duration (i.e., 36 hours rather than 48) and a different release time (mid-afternoon vs. early morning) in Whitmore et al. (2007). Finally 5) The team and individual conditions were done as a *within-participant* factor in Whitmore et al. (2007).

The longer fatigue protocol in the current study, along with early morning release, lead to significant fatigue effects exhibited on every kind of task administered. However had this study stopped at about the same point as Whitmore et al. (2007, at about trial 12 for 18-trialed tasks and at trial 6 for 9-trialed tasks), fatigue effects would *not have been evident* on C3STARS (team or individual) nor SynWin. Hence, there is at least a partial replication of the earlier study with respect to C3STARS exhibiting less fatigue than ANAM-type tasks, even if the forms of the functions in this study are not the same as the earlier one. The similar findings for SynWin and C3STARS points to the importance of

[2] Here we have to qualify that learning in the individual-condition of Whitmore et. al. 2007 occurred for one dimension of mission outcome (friendly fighters lost) but was absent on another dimension (hostile fighters killed, see Whitmore et. al. 2007, Table 1 and Figure 2). If the data in Whitmore et. al. 2007 are re-computed as "mission outcome" in the manner of this study, learning would not reach significance (i.e. 35.9 vs. 37.3 for baseline and full-recovery performance, respectively $t(29) = -1.09$.

task properties in predicting the impact of fatigue -- i.e. a non-asymptotic learning history or a dynamic (and less homogeneous) type of task work may be important as fatigue moderators. This shifts some of the emphasis on fatigue moderators away from teams vs. individuals and more toward task characteristics.

However, there is evidence that teamness was an important *negative* moderator of fatigue in the current study, namely the drop off from fatigue being *greater* in the team condition relative to the additive model portrayed in Figure 5. A bigger fatigue impact in the team condition seems reasonable. If my teammate watches my back less well or stops doing any effective work past a certain level of fatigue, then my effective workload (regardless of my subjective fatigue level at that point) may become much higher relative to an individual-C3STARS player. Saying it another way, we specifically engineered the individual-C3STARS condition to have no teammates that could affect the work load of an individual player (i.e., fighting in a lane and minimizing between-lane interaction). We were also hoping that teammates, in the team condition, would be dependent on each other, even though their team architecture had no *resource-dependency* built into it (i.e., everyone had what they needed to do 1/3 of the war). It appears that fighting a common enemy in the same geographical region was enough to create inter-teammate dependence, inasmuch as the greater decay in performance for teams (relative to the individual condition) is evidence for that, or evidence that a team player's productivity depended not just on their own efforts but their teammates' efforts as well. This characteristic of "teamness" may bias team performance to be more reflective of its weakest member. Other things equal, we should expect a functional team organization (i.e., as in the earlier study) to fair more poorly under extreme fatigue. This follows as each player's productivity is even more heavily interdependent on the others in that architecture. The important words in that last expectation are "all other things equal." In particular, for functional-team architectures, greater team coordination (i.e., increased social interaction) could counteract extreme fatigue's effects in the team context. This issue remains an important issue for future research.

Now we get to the second discussion question for this section. Why do tasks like C3STARS and SynWin look alike in their fatigue functions and look different from ANAM and PVT in their fatigue functions? There is a fatigue literature suggesting complex tasks are less sensitive to fatigue (e.g. see Chee and Choo, 2004, and the literature they cite in this regard); however, psychological models for this finding are not prominent (extant?) in the literature. The most common answer we receive (in terms of comments on the earlier Whitmore et. al. study) focuses on differences in task training and task distance to performance asymptote, prior to entering the fatigue protocol. When participants are not trained adequately, they do additional learning during the fatigue protocol which simply negates the decrements owing to fatigue. The theory might be embarrassed by long flat stretches of unchanging performance, were it not that normative fatigue models have within them long flat stretches of constant performance (which we empirically found, see Figure 2). So "learning-negation" of fatigue may just need to affect the decline to the first sleep-deprivation plateau (which few fatigue studies go beyond) in order to be viable theory for why complex tasks degrade more slowly under fatigue.

With the data in hand the best we can do to assess the learning-negation theory is do a median split on baseline (rested) performance and consider the fatigue functions for the groups above and below median. Our logic is that people who are above the median on initial performance are people who are learning the task faster and are more likely to reach asymptote sometime during the protocol and then show declining performance from fatigue. When this analysis is done for SynWin (the task we have the most data on), the time x group interaction is not significant ($F<1$) and both groups show performance declines from fatigue at the same point in the protocol (trial 7). Plots of the two groups are shown in Figure 8. This finding could suggest learning-negation is not the only factor explaining why SynWin looks different from ANAM or PVT. However, the learning-negation theory might still work if nobody reaches asymptote during the protocol, and fatigue simply becomes severe enough that the learning function (which is peculiarly arousing, at least relative to conventional cognitive tasks, like the ANAM) finally shuts off.

When we did a similar analysis with C3STARS we found more evidence for learning negation going away more quickly in the fast learners. In this analysis we segregated participants into 3 tiers of C3STARS talent and looked at the top 2 tiers for differences in onset of fatigue. Looking at the top tiers throws out the bottom 9 C3STARS participants. Figure 9 plots the results. As can be seen faster learners do seem to start a fatigue decline *earlier*, as we might have expected according to a learning-negation hypothesis (i.e., the tier x epoch x time interaction: $F(10,360)=2.00$, $p<.032$). We might adopt the intra-subject z-score metric to further scrutinize the raw data, as this transformation destroys constant slope and mean differences between conditions (e.g. see Footnote 1, Results), but does not destroy a significant group difference for when a fatigue decline starts to occur, should such a difference exist. In fact, the three factor interaction (fast learners decline sooner than slow learners) is about as significant (i.e., weakly significant) in the intra-subject z-score metric (epoch x time x tier, $F(10,360)=1.94$, $p<.039$).

We don't want to discount the idea that significant learning can occur during a fatigue protocol or that assuring a task is trained to asymptote can make fatigue effects more apparent during the protocol. Some results of the *earlier* study clearly support this, and the learning negation hypothesis – i.e., the team C3STAR condition continued to show learning during the fatigue protocol (and of course the team condition had not been trained to asymptote prior to that). Also the individual C3STARS condition showed fatigue effects (in the middle of the protocol) and clearly had asymptotic learning prior to the fatigue protocol. However, there is also evidence *against* the learning-negation hypothesis *in that earlier study,* namely the individual-C3STARS condition rebounding to near baseline levels at the end of the fatigue protocol. Learning-negation does not predict near-baseline performance rebounds of any sort. When we combine findings like that (and findings like Figure 8), perhaps a simple learning-negation hypothesis cannot explain all the cases one might encounter when one observes a complex, dynamic task having longer sustainment under fatigue (relative to ANAM-like and PVT tasks). If such greater sustainment is evident -- in spite of asymptotic training procedures-- this might mean a differing amount of cognitive arousal (not dependent on new learning) in those

tasks. The characteristics of such cognitively arousing tasks, should they exist, should be factored into fatigue models to get a better normative picture of fatigue in the workplace (i.e., characteristics of the job may matter). However, this issue also remains for future research to determine.

CONCLUSIONS

We must admit to some short-comings with respect to executing this study. For the primary objective, we did not maximally record all the relevant aspects of participant's sleep/wake history that could have biased their classification into either the fatigue susceptible or resistant groups. We were somewhat taken by surprise by the diversity of our population sample in terms of their sleep behavior. However, as we discover factors that are important to participants' classification, we believe our solution of re-ranking the residuals of prediction (from those factors) is a rational and effective strategy for correcting the bias. As discussed before, there are some caveats for when correction is advisable.

As we refine our ways of knowing the starting conditions for our study's participants, we will be able to create better "residual-score" type rules that allow us to see how much normative fatigue models like SAFTE fail to fit the data. As an example, we will be trying to assess "phase shift" parameters and not just sleep-length parameters with respect to biasing a participant's classification. A phase shift parameter is a sleep-wake behavior pattern that fixes a person's circadian rhythm across their daily activities. The most important determinant of a person's circadian phase is their habitual time for going to sleep and waking up (Hursh et. al. 2004). Unfortunately, we did not directly query our participants on this sleep parameter. However, on a future scrutiny of this study's data we will be estimating participant's sleep wake cycle from their activity logs for the week prior to the study. If it turns out that aspects of their circadian phase contribute to their classification probability, then we have a procedure for revising our classifications in light of that new information.

On the other hand we can show that the finding of a substantial number (e.g. 20) of "atheoretic" participants, who are flat-functioned or fatigue resistant cannot be "explained" by SAFTE, no matter how extreme the sleep-dynamics between resistant and susceptible groups are posited to be. For instance, if we compare SAFTE predictions for graveyard-shift workers (i.e., people who go to sleep at 0900 and wake up at 1600) to the observed function for the fatigue-resistant group, the latter is still significantly flatter than what SAFTE predicts for the former. Given this finding, the lack of SAFTE fit in Figure 2 is somewhat explainable by consideration of these atheoretic participants having been averaged into the rest of the sample. Another source of lack of fit is the SAFTE predictions plateauing higher than the observed functions in Figure 2. However, the SAFTE predictions are based on an 8 hour sleep length parameter and when this is adjusted to 7 hours (which is what our population reports getting), the predicted and observed plateaus coincide. Figure 3 also suggests that the most fatigue-susceptible group would fit the SAFTE function the best (i.e., better than the non-displayed middle group, although we haven't assessed this). In summary, given the observations of this

paragraph, we agree with Von Dongen, Baynard, Maislin, and Dinges (2004), that the idea of a normative theory for fatigue impact on tasks, while very useful as a predictive heuristic, still has a lot of new explanatory ground to cover, before quantitative predictions can be made accurately.

The secondary objective of the current study, which is an assessment of fatigue under more complex and socially involved conditions, was also compromised by not enough training to reach a performance asymptote. However, tasks, such as C3STARS, are arguably not trained to asymptote under real-world conditions, as well, and we should point out that knowing the impact of fatigue on tasks, is not something that should *only* be considered for well-learned tasks! Fatigue happens in the context of novel tasks as well, and the fact that learning negates fatigue (as one possible explanation of the aberrant fatigue functions for C3STARS) has potential research interest.

Tasks, like C3STARS, are important for instilling in the participant a work-like atmosphere that could lend more credibility and resilience to the performance of cognitive assessment tasks. An interesting and open question is to what extent fatigue impact on ANAM and PVT would have been the same in a context, not like ours, but more similar to the typical fatigue study (e.g. all tasks resemble our selected ANAM tests and the PVT). A possible intuition might be that there would have been a significant difference (e.g. ANAM and PVT performance over the protocol would be worse under more typical fatigue study conditions). Empirical support for this intuition could be important (and would also help explain SAFTE misfit), given normative theories of fatigue impact, such as SAFTE, have no theoretical mechanism for expecting a difference between testing conditions (e.g. arousal mechanisms).

Finally, we comment on some "metric" issues, or how best to measure fatigue impact. These issues were emergent in many of the analyses of the current study's data. As we have already covered metrics comparing people for fatigue impact, we'll concentrate on issues comparing fatigue impact across situations. We note that the intra-subject z-score transformation proved useful, in a variety of situations, which leads us to believe that it may be an equally valid starting point for the modeling of fatigue impact on task performance, when compared to alternative metrics. As an example of an alternative metric, one often sees fatigued performance divided by baseline performance, where the baseline performance is provided either by the same subjects or by a control group (e.g. Hursh, et. al. 2004, Figure 6). This kind of re-norming of the data, in preparation to comparing the data to theoretical predictions, is highly related to the percent-change metric we used (initially) to rank participants on fatigue-susceptibility.

But how appropriate is a percent-change metric for modeling? We note, as a basis for doubt, that percent-change metrics has some dependence on task or task-measurement scales. Consider a percent change on a reaction time task that increases from a baseline of 1000 msec in 500 msec increments for each succeeding fatigue-measurement point (on average) and another test of 100 questions that declines with 10 incremental errors for (on average) for the same points. Assuming both effects were equally significant in an ANOVA, the 50% and 10% rates of change with fatigue suggests that the reaction time

task is more sensitive to fatigue. However, putting these two testing situations (i.e., their observed averages) in intra-subject z-score units would show equal sensitivity to fatigue. In summary, both metrics can't be right for modeling purposes, and given the z-metric seems more independent of the task (at least in this example), it may be a more suitable, as an upfront performance metric, to map onto theories of fatigue impact, like SAFTE, which express fatigue impact in task-invariant terms, like cognitive effectiveness.

Figure 1. Fatigue functions (performance by hour into the protocol) for raw data on the primary objective tasks. Error bars show a range of 4 std error of the mean.

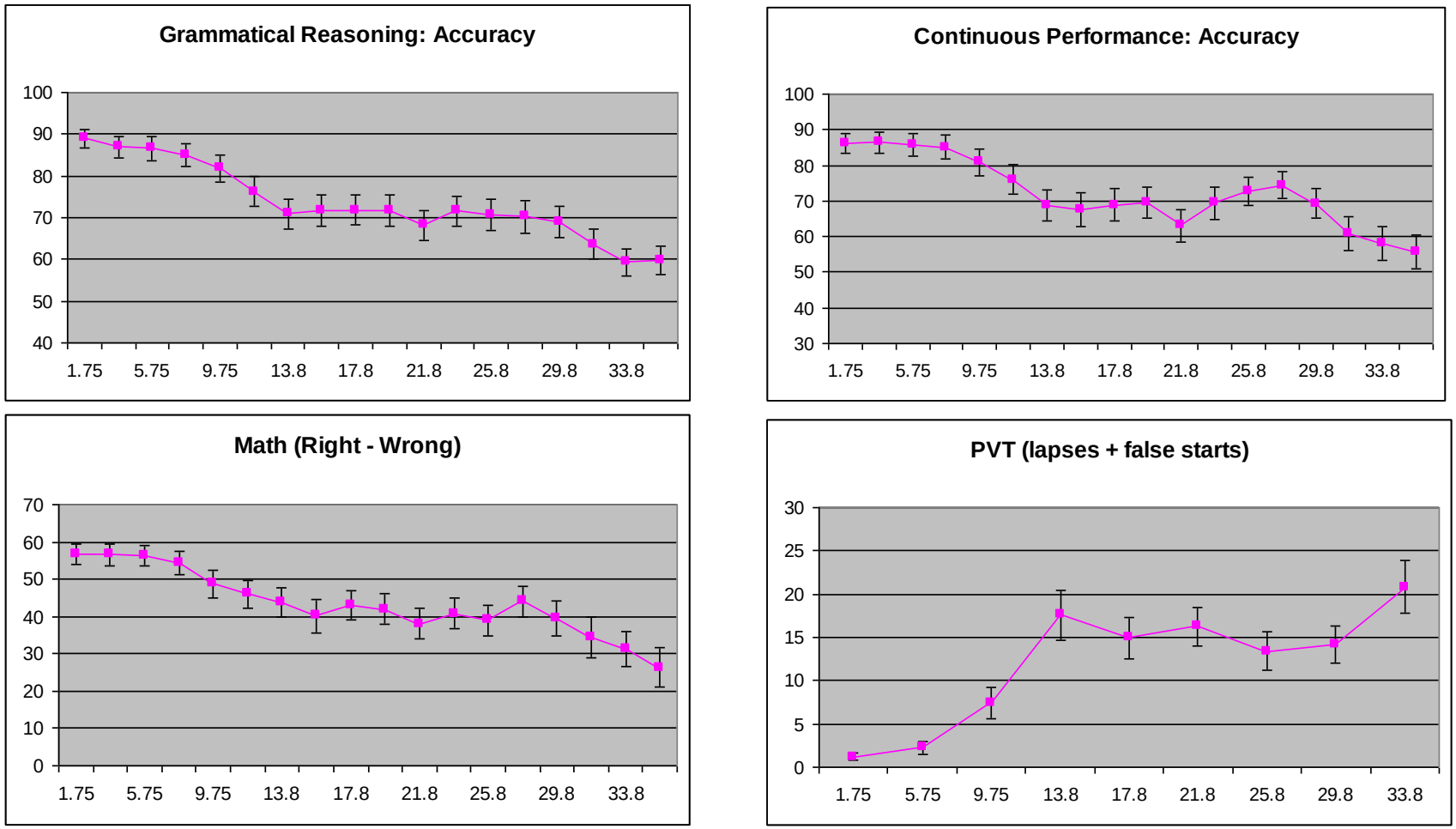


Figure 2. Fatigue functions for primary objective tasks normalized and compared to normalized SAFTE predictions. See text for details.

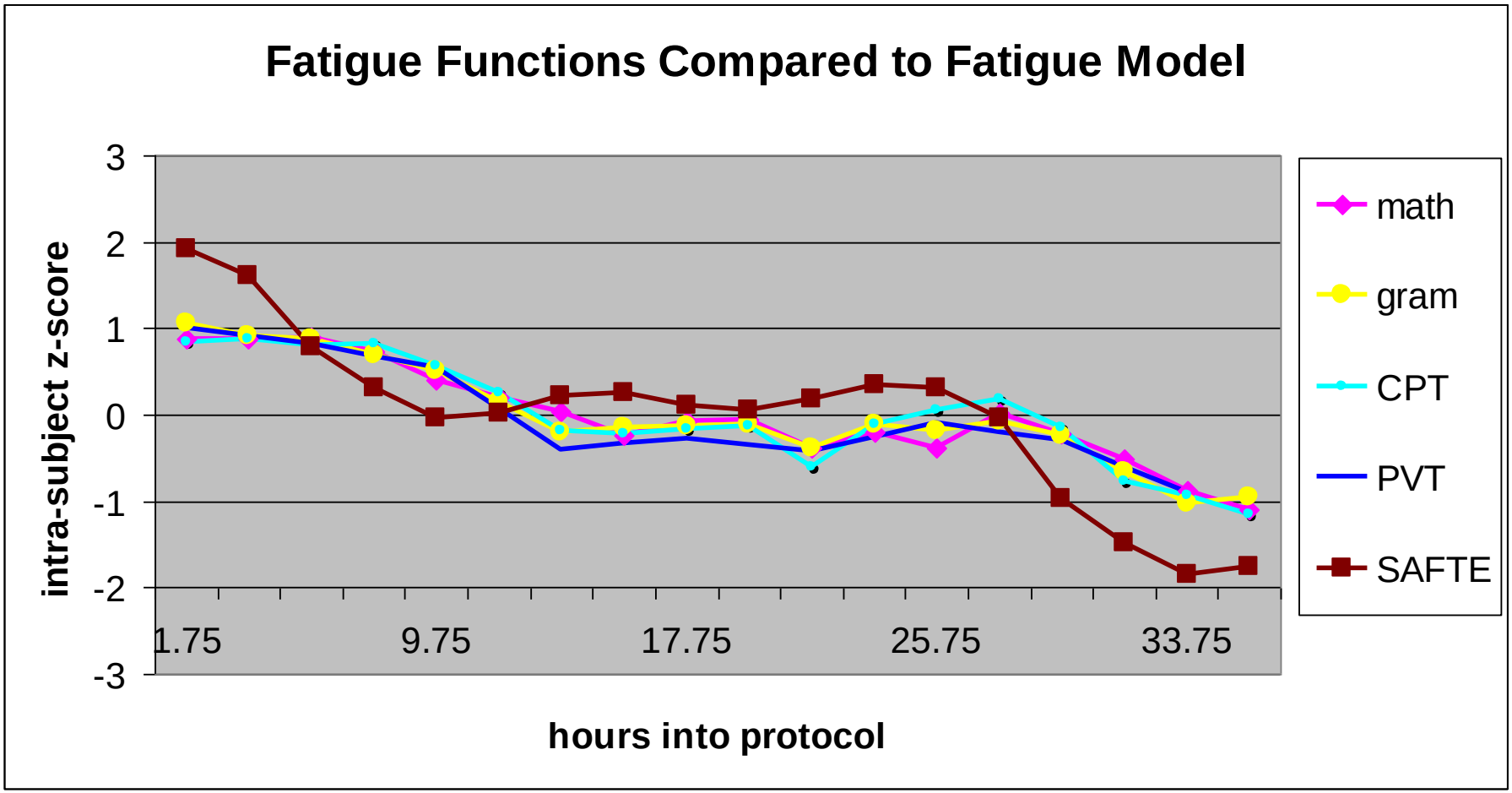


Figure 3. Fatigue plots for fatigue susceptible vs. fatigue resistant based on percent-change rule (left-side) and the residual-score rule / Weekday Sleep covariate (right-side). Error bars show a range of 4 std error of the mean. Fatigue susceptible plots decline more prominently and lay beneath fatigue resistant plots (with the exception of PVT where large scores indicate poor performance)...

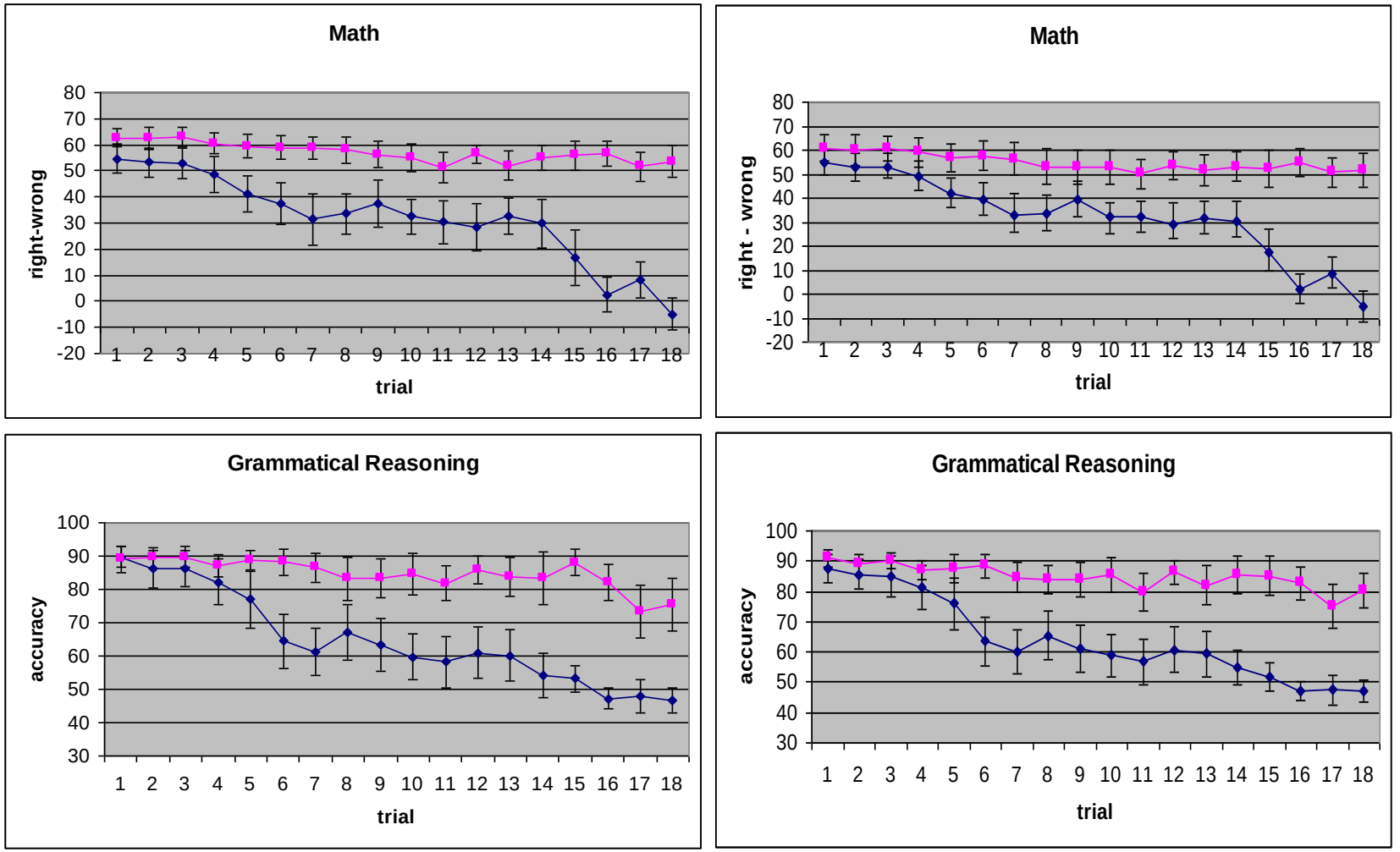


Figure 3 (continued). Fatigue functions for fatigue susceptible vs fatigue resistant based on percent-change rule (left) and the residual-score rule / Weekday Sleep covariate (right-side). Error bars show a range of 4 std error of the mean.

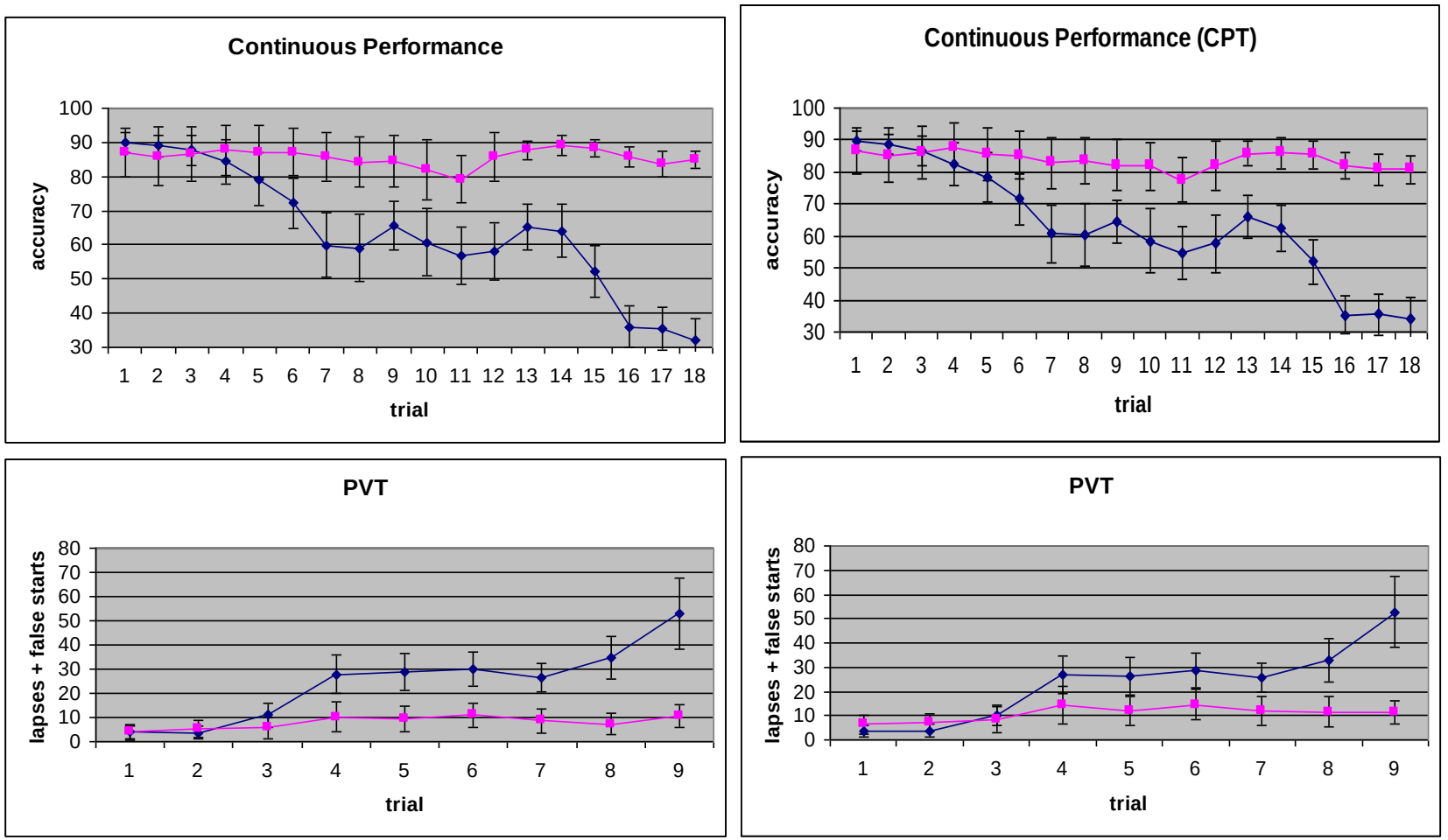


Figure 4. Fatigue functions (performance x trial) for C3STARS for teams (blue-imputed: n=12, pink-full-data-sample: n=5) and individuals (yellow-full-data sample: n=44).

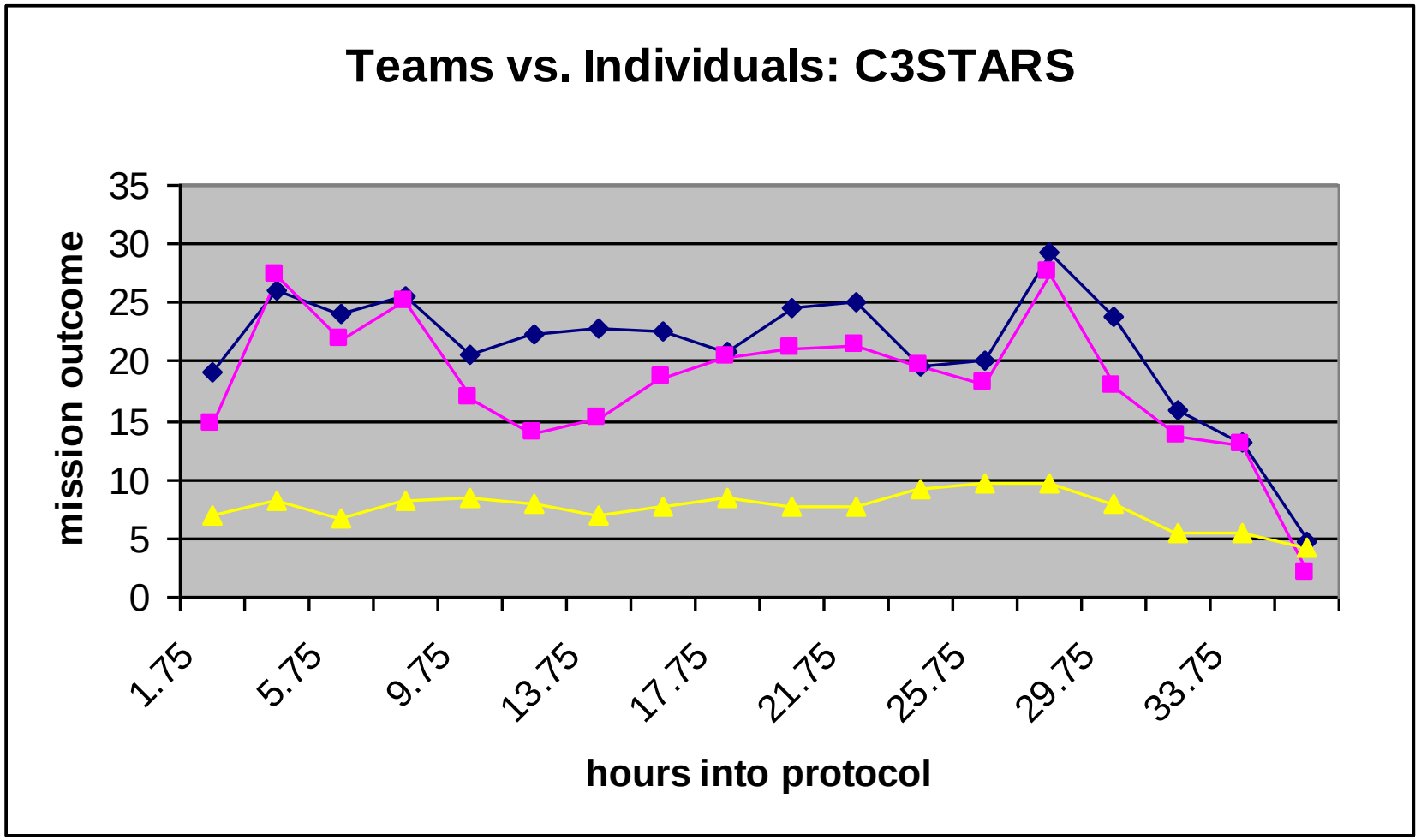
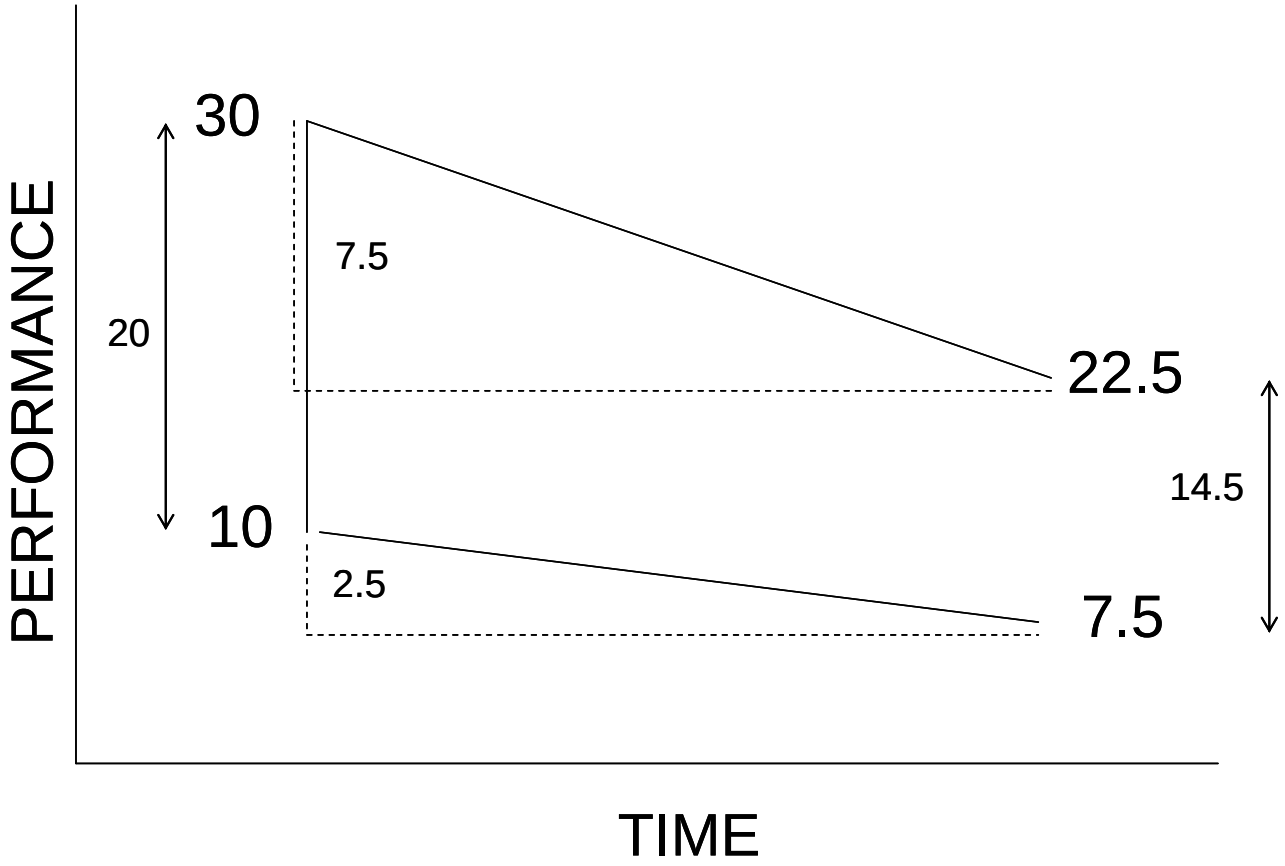


Figure 5. A conceptual schematic showing a significant Team/Individual x Epoch interaction in the raw data is not enough to show differential impacts of fatigue on teams and individuals.



This figure shows two performance points, early (i.e., rested) and late (i.e., fatigued), for both a team and an individual. Each individual can do 10 units of work in a rested trial period and when fatigued they can do 75% of their rested performance. Assuming individuals work independently and additively in the Team condition, the expected means show a significant Team/Individual x Epoch interaction will occur, even while assuming fatigue affects all participants in all conditions the same way.

Figure 6. Figure 4 re-plotted in the intra-subject z-scale. To reduce clutter error bars (4 standard-error-of-the-mean range) are only shown for the first 4 and last 4 points of the protocol and only for the individual (yellow) and imputed (blue) team functions.

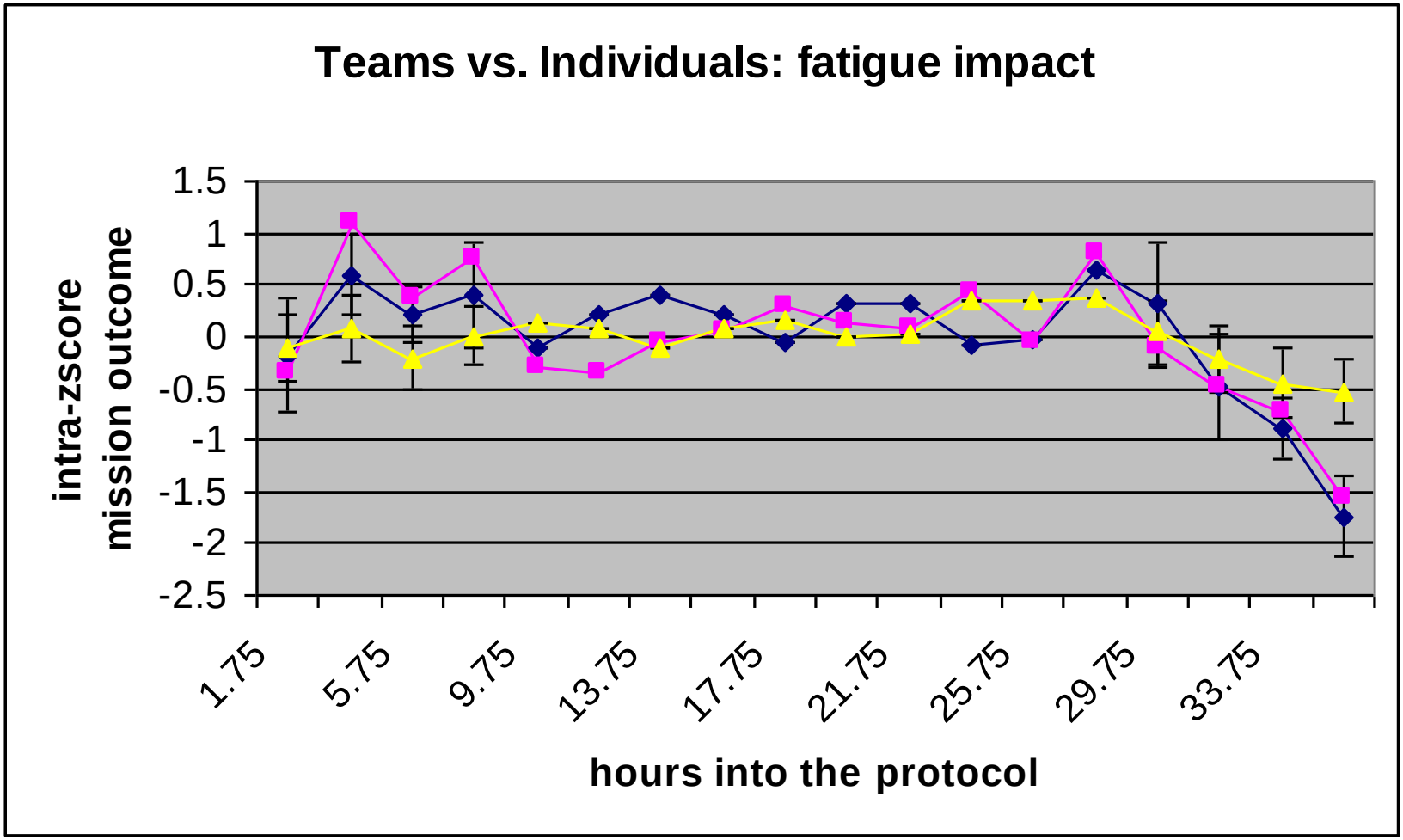


Figure 7. C3STARS Individual condition (C3_I), SynWin, and PVT fatigue functions normalized by within-participant variability. Error bars (4 s.e. of the mean) drawn for C3_I and SynWin only.

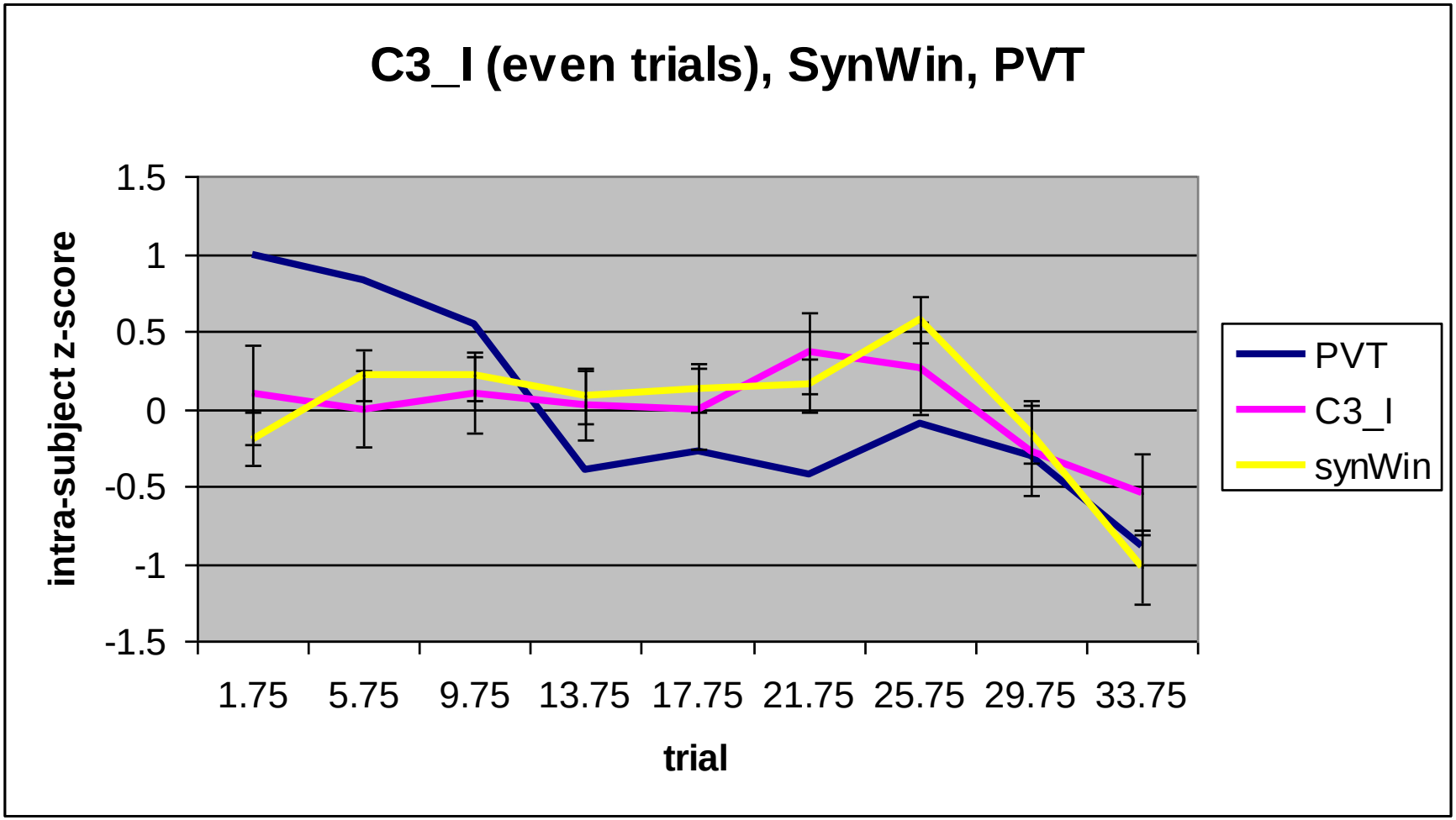


Figure 8. Fatigue functions on the Synthetic Work for Windows Test split by median on the baseline. See Discussion (Secondary Objective) for details.

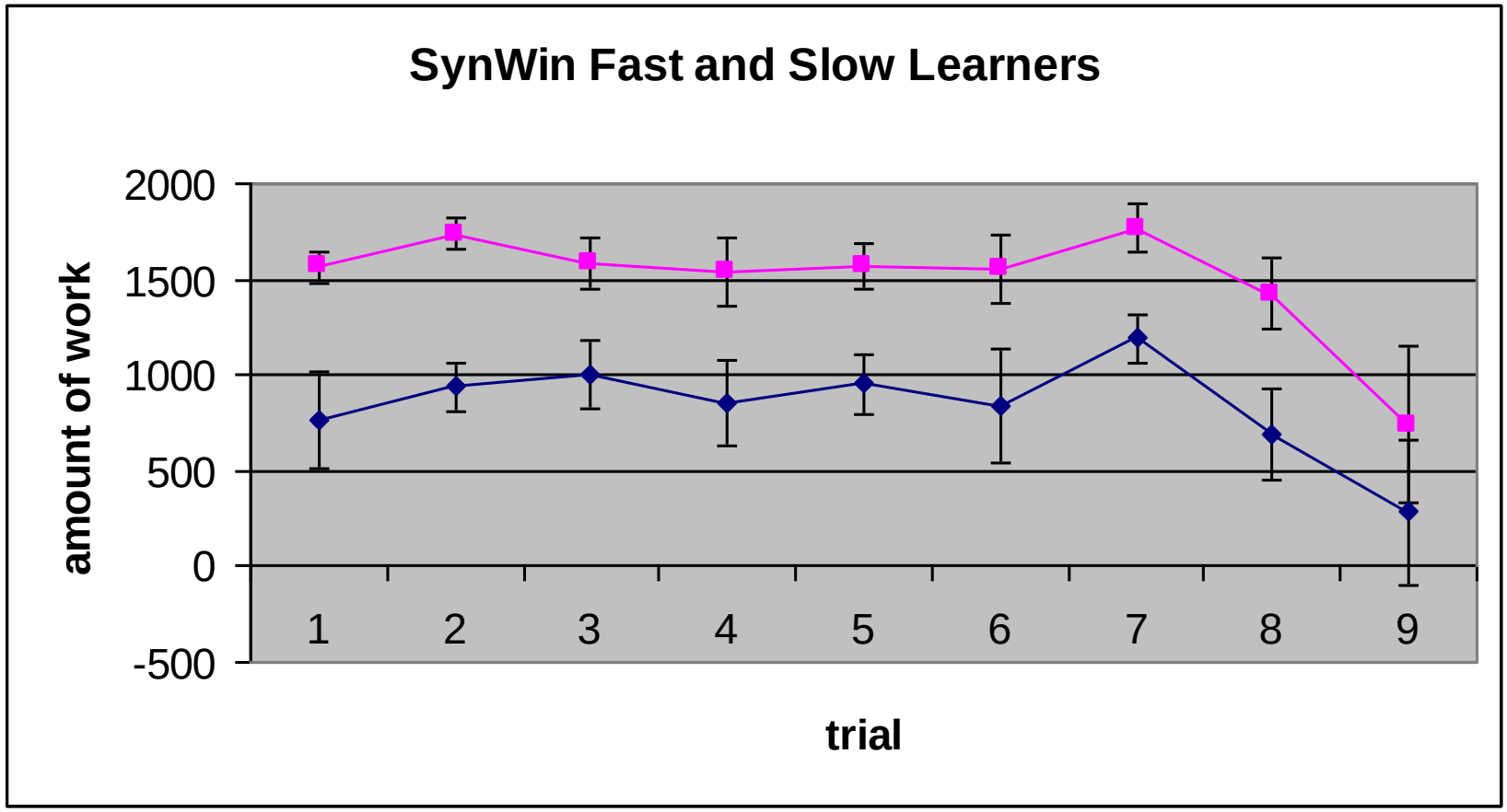
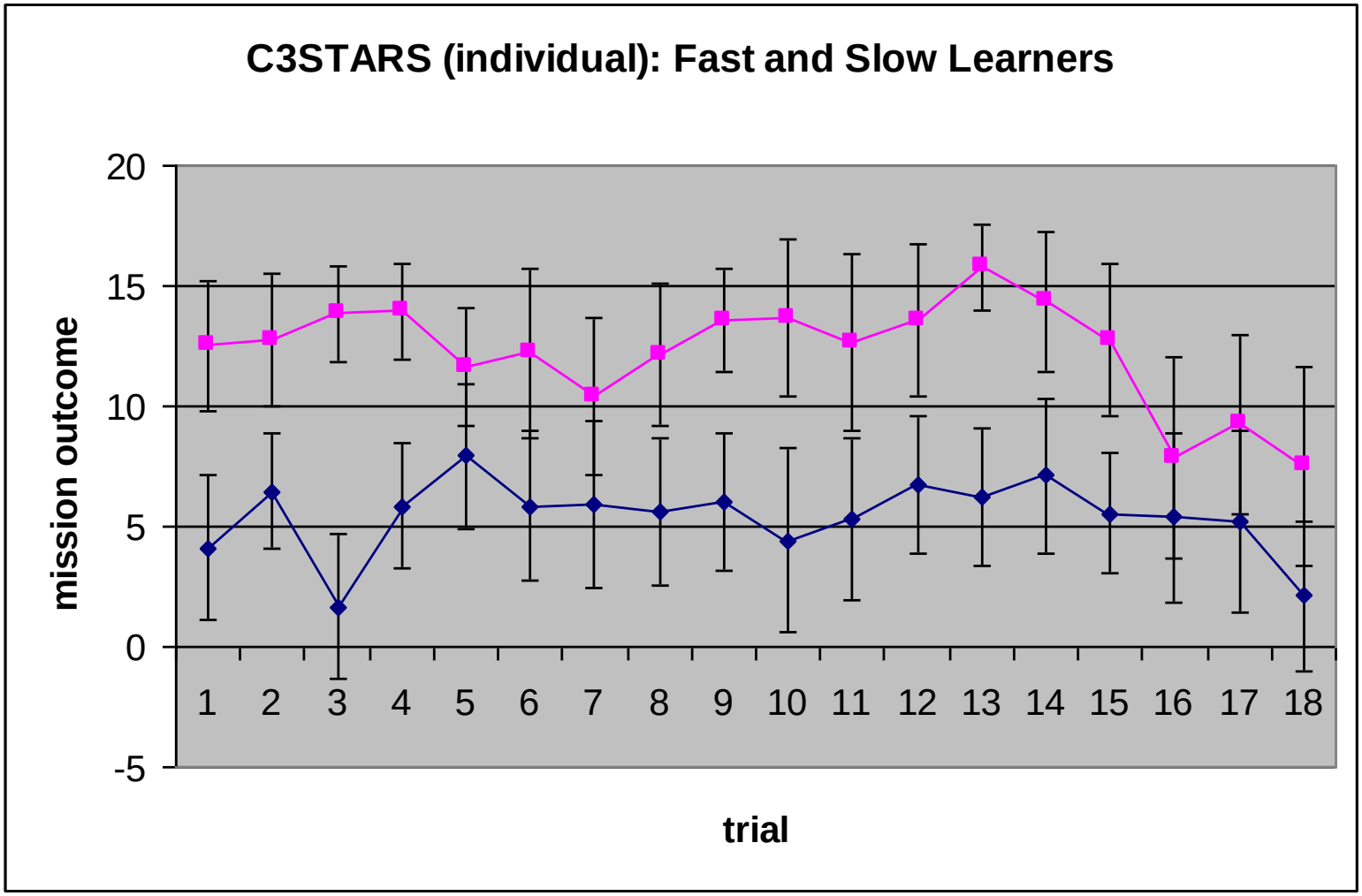


Figure 9. Fatigue functions on C3STARS-individual split by median on the baseline (after the 9 lowest performers are removed). See Discussion (Secondary Objective) for details.



REFERENCES

Barnes, C., & Hollenbach, J. R. (in press). Sleep deprivation and decision-making teams: Burning the midnight oil or playing with fire? *Academy of Management Review*.

Bonnet, M. H. (2000). Sleep deprivation. In M. Kryger, T. Roth, & W. Dement (Eds.), *Principles and practices of sleep medicine* (pp. 53-68). Philadelphia, PA: W. B. Saunders Company.

Caldwell, J. A., Mu, Q., Smith, J. K., Mishory, A., Caldwell, J. L., Peters, G., Brown, D. L., & George, M. S. (2005). Are Individual Differences in Fatigue Vulnerability Related to Baseline Differences in Cortical Activation? *Behavioral Neuroscience*, 119(3), 694-707.

Chee, M. W. L, & Choo, W. C. (2004). Functional Imaging of Working Memory after 24 Hours of Total Sleep Deprivation. *The Journal of Neuroscience*, 24(19), 4560-4567

Cohen, M. S. (1993). The naturalistic basis of decision biases. In G. Klein, J. Orasanu, R. Calderwood, & C. Zsambok (Eds.), *Decision making in action: Models and methods* (pp 51-99). Norwood, NJ: Ablex Publishing Corporation.

Costa, P. T., Jr., & McCrae, R. R., (1992). *Revised NEO Personality Inventory and new five-factor inventory: Professional manual.* Odessa, FL: Psychological Assessment Resources.

Dinges, D., Pack, F., Williams, K., Gillen, K. A., Powell, J. W., Ott, G. E., Aptowicz, C., & Pack, A. I. (1997). Cumulative sleepiness, mood disturbance, and psychomotor vigilance performance decrements during a week of sleep restricted to 4-5 hours per night. *Sleep*, 20, 267–77.

Dorrian, J., Roach, G. D., Fletcher A., & Dawson, D. (2007). Simulated train driving: Fatigue, self-awareness and cognitive disengagement. *Applied Ergonomics*, 38, 155–166.

Gunzelmann, G., Gluck, K. A., Kershner, J., Van Dongen, H. P. A., & Dinges, D. F. (2007). Understanding decrements in knowledge access resulting from increased fatigue. In The 29th Annual Conference of the Cognitive Science Society. Nashville, Tennessee, USA.

Harrison, Y., & Horne, J. A. (2000). The impact of sleep deprivation on decision making: A review. *Journal of Experimental Psychology: Applied*, 6, 236-249.

Hursh, S. R. (1998). *Modeling sleep and performance within the integrated unit simulation system (IUSS)* (Report No. Natick/TR-98/026L). Natick, MA: United States Army Soldier Systems Command; Natick Research, Development and Engineering Center.

Hursh, S. R., Redmond, D. P., Johnson, M. L., Thorne, D.R., Belenky, G., Balkin, T.J., Storm, W. F., Miller, J. C., & Eddy, D. R. (2004) Fatigue Models for Applied Research in Warfighting. *Aviation Space and Environmental Medicine*, 75(3), 44-53.

Klein, G. A. (1993). A recognition-primed decision (RPD) model of rapid decision making. In G. Klein, J. Orasanu, R. Calderwood, & C. Zsambok (Eds.), *Decision making in action: Models and methods* (pp. 139-147). Norwood, NJ: Ablex Publishing Corporation.

Kryger, M., Roth, T., & Dement, W. (2000). *Principles and practices of sleep medicine* (3rd ed.). Philadelphia, PA: W. B. Saunders Company.

Kyllonen, P. C., & Christal, R. E. (1190). Reasoning ability is (little more than) working-memory capacity! *Intelligence*, 14, 389-434.

Lieberman, H. R., Tharion, W. J., Shukitt-Hale, B., Speckman, K. L., & Tully, R. (2002). Effects of caffeine, sleep loss, and stress on cognitive performance and mood during US Navy SEAL training. *Psychopharmcology*, 164, 250-261.

Luna, T. D., French, J., & Mitcha, J. L. (1997). A study of USAF air traffic controller shiftwork: Sleep, fatigue, activity, and mood analyses. *Aviation Space and Environmental Medicine*, 68, 18-23.

McNair, D. M., Lorr, M., & Droppleman, L. F. (1981). *Manual for the profile of mood states.* San Diego, CA: Educational and Industrial Testing Service.

Mitchell, T. R., & Beach, L. R. (1990). ...How do I love thee? Let me count ... Toward an understanding of automatic decision making. *Organizational Behavior and Human Decision Processes*, *417,* 1-20.

National Institute for Occupational Safety and Health. (2004). *Overtime and Extended Work Shifts* (Report No. 2004-143). Cincinnati, OH: Author.

National Sleep Foundation. (2005). *Sleep in America poll*. Washington, DC: Author.

Orasanu, J., & Connolly, T. (1993). The reinvention of decision making. In G. Klein, J. Orasanu, R. Calderwood, & C. Zsambok (Eds.), *Decision making in action: Models and methods* (pp. 3-20). Norwood, NJ: Ablex Publishing Corporation.

Orasanu, J., & Salas, E. (1993). Team decision making in complex environments. In G. Klein, J. Orasanu, R. Calderwood, & C. Zsambok (Eds.), *Decision making in action: Models and methods* (pp. 327-345). Norwood, NJ: Ablex Publishing Corporation.

Penetar, D., McCann, U., Thorne, D., Kamimori, G., Galinski, C., Sing, H., Thomas, M. & Belenky, G. (1993). Caffeine reversal of sleep deprivation effects on alertness and mood. *Psychopharmacology*, 112, 359-365.

Pilcher, J. J., & Huffcutt, A. I. (1996). Effects of sleep deprivation on performance: A meta-analysis. *Sleep*, 19, 318-326.

Presidential Commission on the Space Shuttle Challenger Accident. (1986). Washington, DC: U.S. Government Printing Office.

Rasmussen, J. (1993). Deciding and doing: Decision making in natural contexts. In G. Klein, J. Orasanu, R. Calderwood, & C. Zsambok (Eds.), *Decision making in action: Models and methods* (pp. 158-171). Norwood, NJ: Ablex Publishing Corporation.

Reeves, D., Winter, K., Kane, R., Elsmore, T., & Bleiberg, J. (2001). *ANAM 2001 user's manual* (Special Report NCRF-SR-2001-1). San Diego, CA: National Cognitive Recovery Foundation.

Siegel, J. M. (2005). Clues to the functions of mammalian sleep. *Nature*, 437, 1264-1271.

Van Dongen, H. P. A. (2006). Shiftwork and inter-individual differences in sleep and sleepiness. *Chronobiology International*, 23(6), 1139-114.

Van Dongen, H.P.A, Baynard, M.D., Maislin, G., Dinges, D.F. (2004). Systematic interindividual differences in neurobehavioral impairment from sleep loss: evidence of trait-like differential vulnerability. Sleep, 27, 3, 423-433.

Weaver, J. L., Bowers, C. A., & Salas, E. (2001). Stress and teams: Performance effects and interventions. In P. A. Hancock & P. A. Desmond (Eds.), *Stress, workload, and fatigue* (pp. 83-106). Mahwah, NJ: Lawrence Erlbaum Associates.

Whitmore, J., Chaiken, S., Fischer, J., Harrison, R., & Harville, D. (2007). Sleep loss and complex team performance. In D. De Waard, G. R. J. Hockey, P. Nickel, & K. A. Brookhuis (Eds.), *Human Factors Issues in Complex System Performance* (pp. 55-66). Shaker Publishing: Maastricht, The Netherlands.

Woltz, D.J. (1988). An investigation of the role of working memory in procedural skill acquisition. *Journal of Experimental Psychology: General*, 117, 319-33